JP3.8 RETROSPECTIVE VERIFICATION OF ENSEMBLE STREAMFLOW PREDICTION (ESP): A CASE STUDY

Shuzheng Cong*, John Schaake and Edwin Welles
Hydrology Laboratory, Office of Hydrologic Development
National Weather Service, NOAA, Silver Spring, Maryland

1. INTRODUCTION

The goal of Ensemble Streamflow Prediction (ESP) verification is to give users, forecasters and forecast system managers information to understand the strengths and weaknesses of ESP forecasts. Users may be interested in forecasts for specific locations. Forecasters may be interested in recent forecasts over their area of responsibility. Forecast system managers may want to see evidence that programmatic decisions are leading to system improvements.

Verification of ESP requires large sample sizes to obtain reliable verification statistics. This means that verification of predictions for individual forecast locations and for specific forecast situations requires a retrospective verification approach. This sample size requirement implies that verification of probabilistic forecasts for a current year are possible only for many forecast points over a large area. But such current-year verification statistics may not meet user needs for specific locations. Using the NWS/OHD Extended Streamflow Prediction Verification System (ESPVS), (Riverside Technology,Inc.1999), a retrospective verification study of ESP for the site of Bayard (BAYI4), IOWA on the Raccoon River, a tributary of Des Moines River, was undertaken.

Wilks (1995) has proposed a framework of attributes of forecast verification statistics. These attributes define different facets of forecast accuracy. Two attributes of special interest in this study are *reliability* and *resolution*. Reliability measures show how well forecasts are calibrated. Resolution measures how well observed events agree with calibrated forecasts.

The procedure of ESP has been explained by Day (1985). The essence of ESP is to predict the ensemble of future streamflow hydrographs that would occur given the current initial conditions and an ensemble of future forcing inputs (e.g. precipitation, temperature and potential evaporation).

Retrospective verification requires an archive of meteorological forecasts as well as an archive of observed streamflow. Because the main objective of this study is to introduce a suggested approach to ESP verification, a simplified approach is taken to the future forcing inputs. It will be discussed below how to deal with the meteorological forecast archive requirement in the operational implementation of the proposed procedures. This study, however, assumes that the past climatology is representative of future forcing. Historical time series of precipitation and temperature are used for the required ensemble forcing. ESP generates a trace of future streamflow that is an estimate of what would have occurred in the past if the initial conditions had been the same as they are at present. Each year of past data is a member of the ESP.

2. RETROSPECTIVE ESP

Retrospective ensemble streamflow predictions are produced using the ESPVS for a number of initial forecast times and different forecast periods for the BAYI4 forecast point. For the initial conditions at a given initial time in each historical year, a set of streamflow hydrographs, as illustrated in Figure 1 are generated. Each hydrograph is the result of different meteorological forcing. In this case, in a different historical year.

Each streamflow hydrograph is analyzed to produce a single forecast value such as the streamflow volume during the 30 day period following the initial condition.

The information that is contained in a set of retrospective ESP forecasts from the ESPVS is illustrated in Table 1. The forecasts are for the volume of streamflow for the next 30 days starting from initial conditions on March 15. The units of the volume data values are cubic meters per second - days.

Each row in Table 1 corresponds to an ESP forecast beginning on March 15 of a different year, the first being 1951; the last, 1990, for a total of 40 ESP forecasts. Each ESP forecast in Table 1 has

---

* Corresponding author address: Shuzheng Cong, NWS/NOAA, Office of Hydrologic Development, OHD12, 1325 East_West Highway, Silver Spring,MD 20910; e-mail: shuzheng.cong@noaa.gov

	ENSEMBLE MEMBER YEAR											ENS	ENS
	1951	1952	1953	1954	1955	1956	1957	1958	1959	... 1990	OBS	AVG	STD
1951	304.2	423.2	247.6	143.9	148.3	68.3	84.3	59.3	360.1	... 65.3	365.7	209.3	120.9
1952	429.5	567.4	367.4	251.5	253.1	131.5	172.7	131.7	489.5	... 132.6	604.0	312.7	137.5
1953	261.1	381.2	207.7	104.1	104.0	29.2	41.5	29.4	315.7	... 30.3	201.9	167.8	120.0
1954	76.6	106.5	58.8	29.6	33.2	10.6	19.0	10.8	89.9	... 11.8	32.9	52.5	35.7
1955	203.0	316.3	152.6	54.2	53.5	10.9	21.3	11.1	255.2	... 12.0	53.5	124.8	104.5
1956	43.6	64.0	32.5	11.1	12.3	3.6	4.8	3.8	53.2	... 4.8	3.2	27.6	22.9
1957	80.7	165.0	59.0	27.9	28.4	4.7	12.7	4.9	109.3	... 5.8	52.5	58.7	51.2
1958	444.3	591.2	382.7	262.4	264.7	129.9	180.6	130.0	509.7	... 130.9	120.3	335.6	146.9
1959	91.6	147.9	71.1	39.5	43.0	16.4	26.6	16.5	106.9	... 17.5	134.9	64.5	43.7
.....													
1990	529.0	668.3	465.8	348.4	349.9	218.3	268.0	218.5	589.0	... 219.4	192.1	420.5	142.9
OBS	365.7	604.0	201.9	32.9	53.5	3.2	52.5	120.3	134.9	... 192.1			
AVG	353.6	463.6	308.4	220.7	222.1	125.3	162.0	119.2	410.4	... 123.5			
STD	237.3	278.2	225.1	193.0	195.1	143.1	169.6	141.8	253.6	... 149.7			

Table 1- Example ESP Results

40 members. Each member is the volume of streamflow that would have occurred in the forecast year as a result of both the initial conditions in the forecast year and the meteorological forcing that occurred in the year associated with that member. Member values for each ESP forecast are in the same row and vary across the columns of Table 1. The year associated with the forcing for each member is given as a column label at the top of each column. The initial condition dates are given as the row labels for each ESP forecast. Note that the data displayed in Table 1 is compressed so that only some of the data values for the full 1951-1990 period actually appear in the Table.

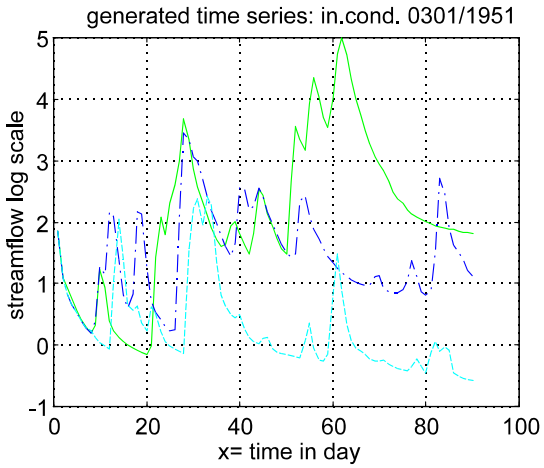

Figure 1 - Example ESP time series, in CMS

Statistics of the data in Table 1 are given in the margins of the Table. The observed streamflow volume is given for each year as well.

The diagonal elements of Table 1 are the model simulated volumes for the same year as the observed meteorological forcing. Figure 2 shows that these simulated volumes (1951-1990) agree very well with the observed volumes. The correlation coefficient between the simulated and observed volumes is 0.97. There appears to be a slight tendency for the largest observed volumes to be greater than simulated.

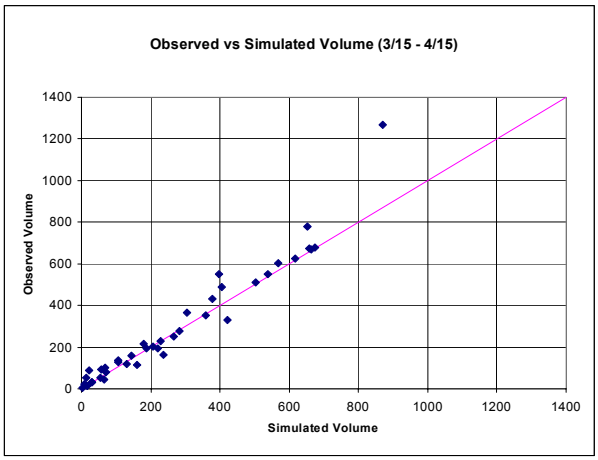

Figure 2 - Simulated vs Observed volume

Each row of Table 1 gives the retrospective ESP forecast that would have been made in the year corresponding to that row. The statistics at the end of each row are the ensemble mean and standard deviation of the ESP forecasts. Figure 3 compares

the ensemble mean with the observed value for each year. The correlation coefficient is 0.82. There is a slight tendency for the largest observed volumes to be greater than the corresponding ensemble means. Note that the range of ensemble mean volumes is almost as great as the range of observed volumes. The scatter of points in Figure 3 is caused by uncertainty in the future meteorological forcing. The strong tendency for the observed volumes to vary with the ensemble mean occurs because the observed volumes are more sensitive to the initial conditions than to the meteorological forcing.

Figure 4 compares the ensemble standard deviation to the ensemble mean volume. Note that for small values of the ensemble mean the ensemble standard deviation increases roughly in proportion to the ensemble mean. But at larger values of the ensemble mean, the standard deviation reaches a plateau.

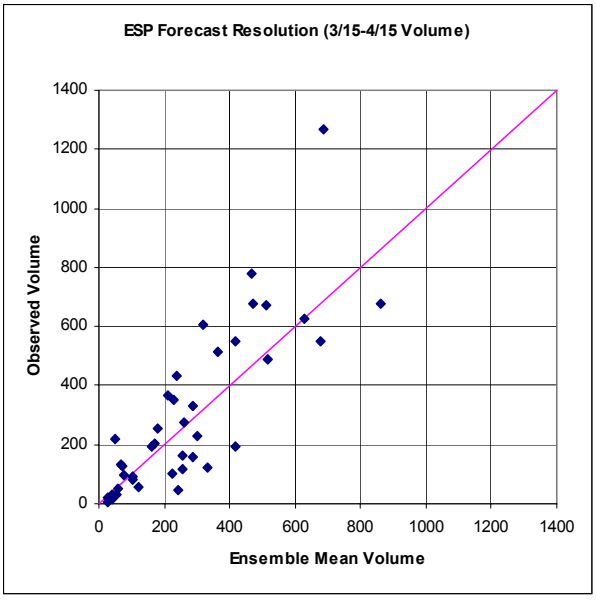

Figure 3 - ESP ensemble mean vs observed volume

Each column of Table 1 shows how the streamflow volume for a given year's forcing depends on initial conditions. Statistics at the bottom of each column give the mean and standard deviation of the estimated streamflow volume that would have occurred for the given year's forcing if the initial conditions had been the same as they were in different years corresponding to the different rows.

Figure 5 compares the mean simulated volume for each column with the observed volume for the year corresponding to the forcing for that year. Note that it is the initial conditions that vary within a given column while the forcing is the same for every value in the column. The scatter of points in Figure 5 occurs because the observed volume is very sensitive to the initial conditions and the column means are not as sensitive to the meteorological forcing. The correlation coefficient is 0.45.

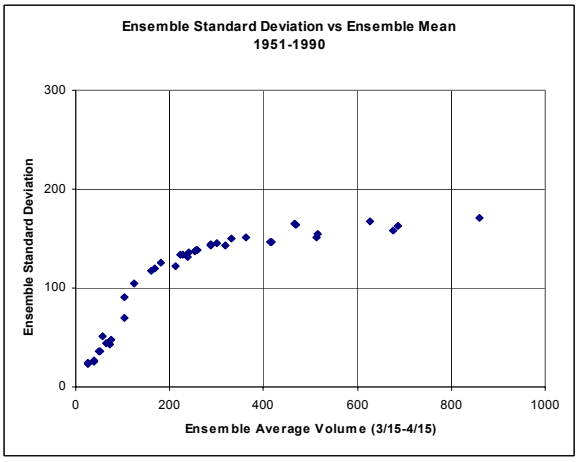

Figure 4 - Ensemble standard deviation vs ensemble mean volume

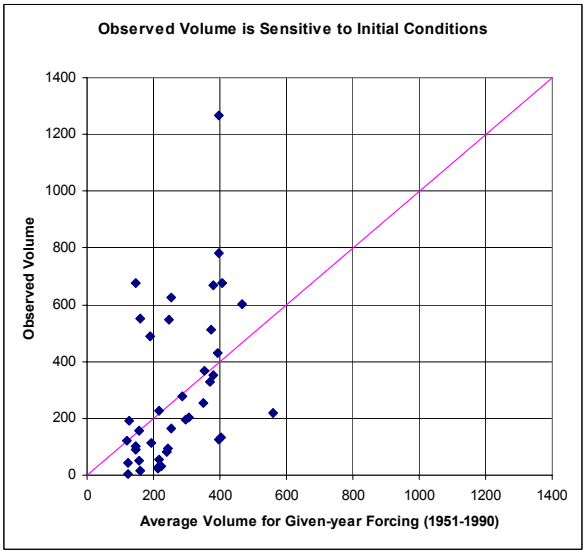

Figure 5 - Observed volume is sensitive to initial conditions

The cumulative probability distribution functions of the observed, simulated and ensemble member volumes are presented in Figure 6. These distributions are very similar but there is a tendency for the largest observed values to be greater than the simulated and the ensemble member values. The distribution of ensemble member values is very close to the distribution of simulated values.

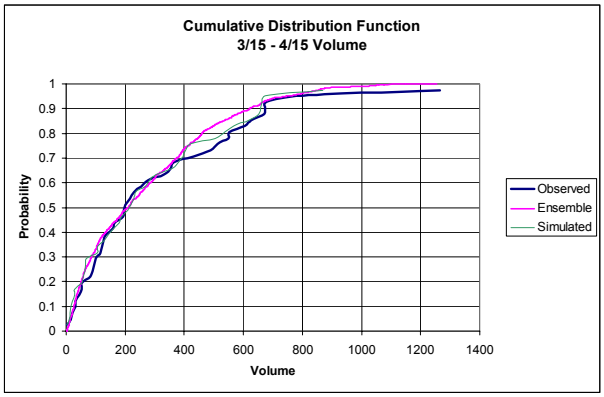

Figure 6 - Cumulative distribution functions of observed, simulated and ensemble member streamflow volumes

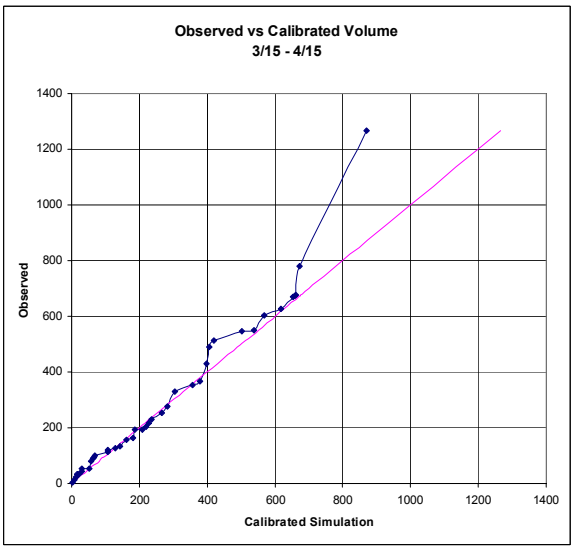

Figure 7 - Observed vs simulated cumulative distribution volumes

The tendency for the model to underestimate larger than average volumes is illustrated in Figure 7 where corresponding pairs are shown of observed and simulated volumes for the same probability level of the marginal distributions shown in Figure 7.

Table 1 can be sorted so that the ensemble mean values increase and so that the column means also increase. Sorting by the ensemble means causes the ensembles to be sorted from dry to wet initial conditions rather than chronologically. Sorting by the column totals causes the ensemble members to be sorted from dry to wet years of meteorological forcing rather than chronologically. The effect of this sorting is to produce a streamflow volume response surface that is relatively smooth. Contours of this response surface are shown in Figure 8.

Left to right cross-sections of the surface in Figure 8 correspond to an ESP for a given year. This cross-section shows how different meteorological conditions affect the distribution of streamflow volumes for given initial conditions. The initial conditions vary from dry to wet as the cross-section is moved upward with increasing value of the vertical y-axis. The value of the y-axis is an index that points to the year corresponding to the initial conditions that increase from dry to wet.

Cross-sections taken from bottom to top show how initial conditions affect the distribution of streamflow volumes that would be produced by fixed meteorological forcing as initial conditions vary from dry to wet. The corresponding meteorological forcing varies from dry to wet as the position of the cross-section is moved from left to right.

Figure 8 shows that the response surface is much more sensitive (i.e. the gradient of the surface changes more) to initial conditions than to the meteorological forcing. Meteorological forcing has almost no effect for very dry initial conditions. If the contours of this surface were parallel to the horizontal axis, the ESP forecasts would depend only on the initial conditions and there would be no uncertainty in the forecasts. If the contours were parallel to the vertical axis: (I) initial conditions would have no effect; (ii) ESP forecasts would depend only on the meteorological forcing; (iii) the forecasts would be the same every year; (iv) the uncertainty in the forecasts would be the same as the climatological uncertainty of the streamflow volume, and (v) there would be no skill in the forecast. It follows then that the slope of the contours in Figure 8 is a measure of the local skill in the forecast. Horizontal contours have complete skill. Vertical contours have no skill.

Figure 8 can be used to put a current operational ESP forecast into perspective relative to ESP forecasts for other years. The ensemble mean of the current ESP forecast can be used to find the position on the vertical axis corresponding to the forecast location for the current year's forecast. One

could then draw a line across Figure 8 at that location and the distribution of values of the ensemble members of the current ESP forecast would lie along that line and can be visually compared to forecasts for other wetter or drier years.

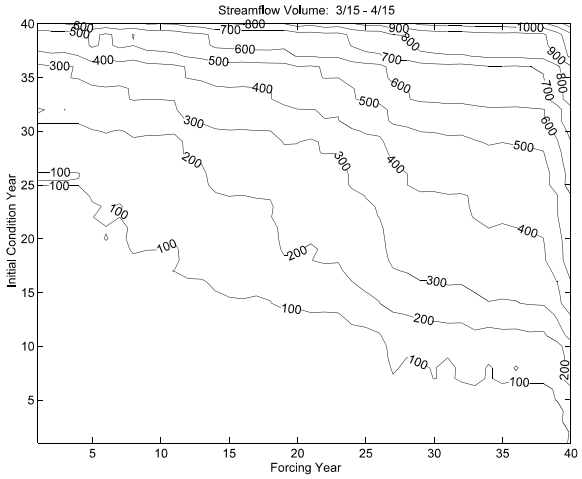

Figure 8 - Retrospective ESP streamflow volume response surface

3. PROPOSED ESP VERIFICATION STATISTICS

Verification statistics are proposed to measure various aspects of the accuracy of ESP forecasts. One objective is to include statistics that may be important to various users and that can be applied to individual forecast points. Another is to suggest statistics that are dimensionless so that they potentially could be aggregated (using an appropriate rule) to provide summary information for forecasters, forecast developers and management. Also it may be possible to form regional aggregates of some of these statistics over large enough areas to get sufficiently large samples of independent events.

3.1 *Ensemble Mean Correlation Coefficient*

A statistic that is a direct measure of forecast resolution is the correlation coefficient between the mean of each ensemble forecast and the observed value. This statistic, by definition, is not affected by biases in the mean or standard deviation of the forecast. It's value is in the interval (-1,1). Zero implies no skill relative to climatology. A value of 1 implies perfect correlation (But the forecast and observed values are not necessarily the same because of systematic biases.)

3.2 *Ensemble Mean Skill Score*

A statistic that is closely related to the correlation coefficient is the Nash-Sutcliffe efficiency statistic (EnsSS). In the verification literature (e.g. Doggett, 1998) this statistic would be called a skill score because the value of the statistic is scaled by the climatological variance of the observations. The statistic is

$$EnsSS = 1 - \frac{(1/n)\Sigma(ESPavg-OBS)^2}{(1/n)\Sigma\,(OBS-OBSavg)^2} \qquad (1)$$

where,

ESPavg = Ensemble mean
OBS = Corresponding observation
OBSavg = Average observation

and the summation is taken over the set of n ESP forecasts for a given starting time and forecast period.

If the forecasts are unbiased (in the mean) and have minimum error variance, then EnsSS is equal to the square of the correlation coefficient. In that case EnsSS is a direct measure of forecast resolution. Because EnsSS is affected by forecast bias, it is a composite measure of resolution and reliability.

3.3 *Measures of Forecast Reliability*

Two direct measures of reliability considered in this study are the relative bias of the ensemble mean forecast (B),

$$B = (1/n)\,\Sigma ESPavg / OBSavg - 1 \qquad (2)$$

and a root mean square error statistic that measures the reliability of the ESP probability forecasts.

The reliability of probability forecasts can be assessed by constructing what is known as a reliability diagram (Wilks(1995)). An example reliability diagram is illustrated in Figure 9 for forecasts from Table 1. The reliability diagram is created as follows. First, each ensemble forecast is used to find the forecast probability of observing a value less than or equal to the observed value for that forecast. Then, these probabilities are sorted in increasing order. If the forecasts were perfectly reliable these probabilities would form a uniform distribution. Therefore, points on the uniform probability distribution are plotted on the forecast probability axis of Figure 9 and the observed relative

frequency is taken from the ESP probability associated with the observation. In the example, there is a tendency for the observations to occur more frequently than expected by chance (because the observations are actually slightly larger than expected by the forecast model). This is consistent with the tendency shown in Figures 2, 3, 6 and 7 for the model to underestimate higher than average streamflow volumes.

The magnitude of the vertical deviation of the observed relative frequency from the 45-degree diagonal in Figure 9 is a measure of reliability (RMSrel). In this study, we propose using the RMS value of this deviation as a measure of reliability of the forecast probabilities.

$$RMSrel = ((1/n) \Sigma ((Fobs - Funiform)^2)^{1/2} \quad (3)$$

A useful graphical tool, in addition to the reliability diagram is the Talagrand diagram. The Talagrand diagram is a plot of the number of times the estimated probability of non-exceedance (or exceedance) of each observation is found to occur in different probability intervals. This is illustrated in Figure 10 for March 15. The integral of the Talagrand diagram is consistent with the reliability diagram. In this case the Talagrand diagram shows a tendency to underestimate the probability of large events occurring.

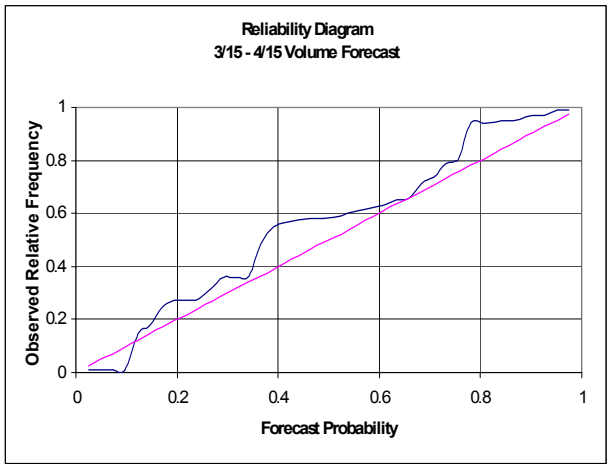

Figure 9 - Example reliability diagram

3.4 *Heidke Skill Score*

Verification of probability forecasts by the NWS Climate Prediction Center is done using the Heidke skill score (HSS). Probability forecasts are made for events to occur in terciles corresponding to below (B) near (N) and above (A) normal relative to the climatological distribution of the observations. Intervals of the forecast variable are defined so there is a 1/3 chance, climatologically, of an event occurring in each tercile. ESP forecasts can be used to make categorical forecasts for below, near or above normal forecasts by assigning the forecast to the tercile with the highest forecast probability.

The Heidke skill score is defined as

$$HSS = (H - E)/(T - E) \quad (4)$$

where

H = number of hits - a hit is defined as the number of times the observed event occurs in the tercile with the highest forecast probability, given that the forecast meets the criteria to be included (see below)

T = number of forecasts - only forecasts where the probability for below or above normal exceeds a threshold level are included.

E = expected number of hits by chance (=1/3 T)

In this study, the threshold probability is set at 1/3. By adjusting this upward, it might be possible to improve the Heidke skill score.

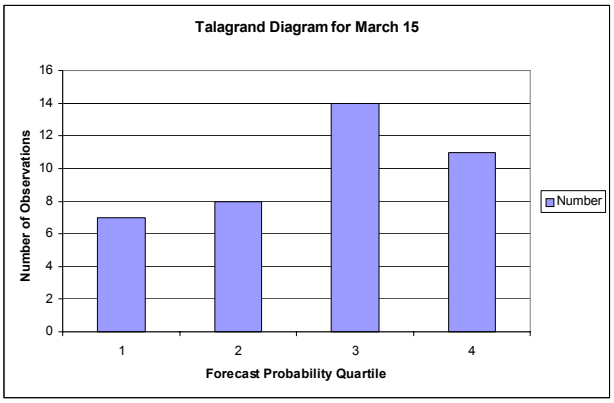

Figure 10 - Talagrand diagram for March 15

CPC uses a probability threshold to classify forecasts as either above (A) or below (B) normal. If neither the above nor below normal probability values exceed the threshold, the forecast is assumed not to be distinguishable from climatology. Therefore, only events where there is predicted to be some skill are included in computing the skill score. Accordingly, it may be possible to raise the tercile probability threshold to identify areas (in the case of CPC) or

times (in the case of retrospective ESP) with higher probability above or below average.

In a highly skillful ESP forecast, it might be possible to distinguish near-normal forecasts from climatology. This might occur for very short range forecasts that are well calibrated and where the forecast is highly dependent on initial conditions. This could occur, for example, during streamflow recessions.

In this study we prefer to maintain the CPC application of the Heidke skill score and use the Brier skill score to explore the relative skill in below, near or above normal categorical forecasts.

3.5 *Brier Skill Score*

The Brier skill score (Doggett, 1998) is defined as

$$BSS = 1 - BS / BSC \quad (5)$$

where BS is the Brier score,

$$BS = (1/n) \Sigma (p_i - I(obs_i))^2 \quad (6)$$

p_i = probability of event i occurring
$I(obs_i)$ = indicator variable (1 if event occurs, else 0)
n = number of events

and BSC is the climatologically expected value of BS,

$$BSC = p * (1-p) \quad (7)$$

where p is the climatological probability of the event. The Bier score is often applied to events that exceed a given threshold. But it can also be applied to categorical events. In this case we consider the three tercile categories used by CPC. Accordingly, BSC = (1/3)*(1 - 1/3)) = 0.2222 for each category.

4. EXAMPLE APPLICATION

The verification statistics proposed above are applied to the forecast point Bayard (BAYI4), IOWA on the Raccoon River, a tributary of Des Moines River. ESP forecasts for this location have been made during the northern plains spring snow-melt period for the past three years. BAYI4 is one of the original forecast points in the NWS AHPS demonstration project for the Des Moines river basin. Retrospective ESP verification is done below for 30-day volume forecasts for three different starting dates: March 1, March 15 and April 1 The retrospective forecast period is 1951-1990. Verification statistics for these forecasts are given in Table 2.

The correlation coefficient and ensemble mean skill score reach their maximum values on March 15. This is when the initial conditions have their maximum average effect on the forecast.

There is a 5 to 10 percent relative bias depending on the forecast starting date. And there is a small RMS bias in the probability forecasts. These systematic biases could be removed by statistical post-processing of the forecasts. If that were done, the ensemble mean skill score would improve slightly as well. But the correlation coefficient would remain unchanged.

The Heidke and Brier skill scores for categorical forecasts show interesting results. Unlike the correlation coefficient and Ensemble skill score, the Heidke skill score continues to increase from March 1 to April 1. We plan to study why that happened. The Brier skill score helps us to understand what part of the ESP probability forecast tends to be the most skillful. In all 3 forecast periods, the portion of the probability forecasts for above or below normal are more skillful that the portion near normal. This happens because more of the individual ensemble probability forecasts tends to be distributed over the near normal interval than over the above or below normal intervals. When the distribution shifts toward above or below normal, the corresponding event tends to happen more often than expected by chance.

Verification Statistic	March 1	March 15	April 1
Ensemble Mean - Obs Correlation	.71	.82	.65
Ensemble Mean Skill Score	.50	.66	.41
Relative Bias	-.05	-.09	-.08
RMS Error of Probability Forecast	.07	.08	.06
Heidke Skill Score	.225	.294	.339
Brier Skill Scores by Tercile	.208 -.068 .500	.355 .078 .564	.517 .093 .255

Table 2 - Example ESP Verification Statistics for 30-day streamflow volume forecasts for Bayard (BAYI4), IOWA on the Raccoon River

5. CONCLUSIONS

Retrospective verification of ESP can provide useful information about ESP performance. Several potentially useful verification statistics are proposed to measure forecast resolution and reliability, both as separate attributes of forecast accuracy and their joint effect on the forecast. A retrospective ESP response surface was introduced as a way of creating a graphical understanding of ESP forecast skill and of placing any given ESP forecast in perspective to ESP forecasts for wetter or drier years.

This study used historical climatological data for meteorological forcing for the retrospective ESP forecasts. If retrospective meteorological forecasts were available for the entire retrospective period they could have been used instead of the climatological data. Of course that would have required a very long archive of ensemble meteorological forecasts and these forecasts would have needed to be integrated into consistent forecast time series for the duration of the ESP forecast time period. In practice, it will be necessary to use more limited forecast archives for forecasts for different durations into the future. This will require long enough archives from a stable forecast operation to accurately estimate the climatological statistical properties of the meteorological forecasts and it will require techniques to simulate equivalent forecasts that could have occurred before the archive began.

This study also brings up a question of how to use the retrospective verification to improve the ESP through statistical post processing. We will discuss this problem in another paper.

6. REFERENCES

Day, G.N., 1985. “Extended Streamflow Forecasting Using NWSRFS”, Journal of Water Resources Planning and Management, ASCE, 111(2), pp157-170

Day, G.N, et al. 1992 “Verification of the National Weather Service Extended Streamflow Prediction Procedure”, Managing Water Resources During Global Change, American Water Resources Association,

Doggett, K., 1998. Glossary of Verification Terms, http://www.sel.noaa.gov/forecast_verification/verif_glossary.html

Riverside Technology, inc. 1999. National Weather Service Extended Streamflow Prediction Verification System (ESPVS) DRAFT User Manual

Wilks, Daniel S., 1995
Statistical Methods in the Atmospheric Sciences. Academic Press, 467pp.