JP3.7 SHORT-RANGE ENSEMBLE PRECIPITATION FORECASTS FOR NWS ADVANCED HYDROLOGIC PREDICTION SERVICES (AHPS): PARAMETER ESTIMATION ISSUES

John Schaake*, Mary Mullusky, Edwin Welles and Limin Wu
Hydrology Laboratory, Office of Hydrologic Development, National Weather Service (NWS), Silver Spring

1. INTRODUCTION

Reliable Ensemble Streamflow Prediction (ESP) requires unbiased ensemble precipitation forecasts as input to hydrologic forecast models. Because meteorological forecast information may contain a variety of different kinds of biases, an ESP preprocessing system is used in NWS River Forecast Center (RFC) operations to remove these biases. These statistical preprocessing techniques have parameters (coefficients and exponents) that must be calibrated. The calibration varies spatially and seasonally throughout an RFC area of responsibility. In mountainous areas the calibration may vary with each hydrologic sub-basin. An historical meteorological forecast archive of forecasts representative of current operational forecasts is needed to provide data to calibrate statistical preprocessor parameters. Typically the length of available archive is quite limited so uncertainty in the parameter estimates may pose important limitations on the skill of hydrologic forecasts. This paper analyzes the effect of archive duration on the accuracy of preprocessor parameter estimates and on the skill and reliability of the adjusted short range ensemble precipitation forecasts

2. NWS SHORT-RANGE PREPROCESSOR AND ENSEMBLE PRECIPITATION GENERATOR

The initial strategy to develop short range (1-5 days) ensemble precipitation forecast applications for AHPS is to synthesize ensemble forecasts from existing deterministic forecasts produced by the Hydrometeorological Prediction Center (HPC) and used in existing RFC forecast operations. Ultimately, the strategy is to apply ensemble forecasts from regional and global ensemble forecast systems operated at the National Centers for Environmental Prediction (NCEP). But hydrologic application of atmospheric ensemble forecasts requires substantial additional research and development (including defining a role for human forecasters to add value) before a reliable operational system can be implemented.

The basic strategy of the simplified short range ensemble precipitation preprocessor is to use a simple statistical model to account for the uncertainty in existing deterministic forecasts. This strategy is being applied to temperature, or other variables, as well as precipitation. Uncertainty in precipitation forecasts is particularly challenging to represent because of the intermittent nature of precipitation and because of the significant scale-dependency of the skill of the forecasts.

An example joint distribution of precipitation forecasts and observed precipitation events is shown in Figure 1. The data in Figure 1 can be used to create a statistical model of the joint distribution of forecasts and observations. This joint distribution can then be used to create a conditional distribution for the observations, given a forecast. And this conditional distribution can be used to create an ensemble of precipitation events by re-scaling observed historical events . This is essentially a Bayesian approach and Bayesian techniques can be used to estimate the model parameters.

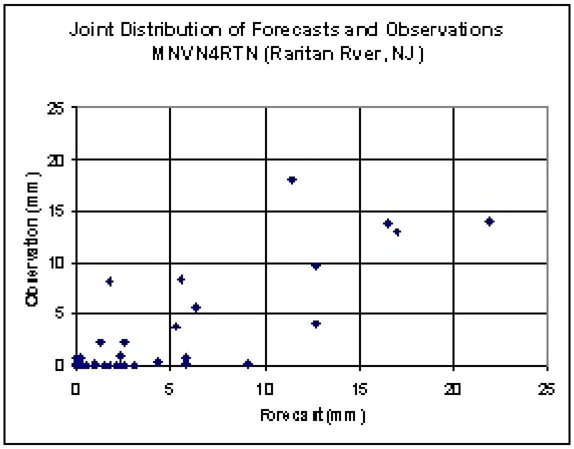

Figure 1 - Joint distribution of forecasts and observations

---

* Corresponding author address: John Schaake, NWS/NOAA, Office of Hydrological Development, OHD12, 1325 East West Highway,
Silver Spring, MD 20910; e-mail: John.schaake@noaa.gov

The joint distribution statistical model has seven parameters that must be estimated from historical data. Most (6) of these can be estimated by considering the marginal distributions of the forecasts and the observations. The marginal distributions represent the climatology of the forecasts and the observations. These marginal distributions can be parameterized. Climatological statistics such as probability of precipitation (POP), mean of the conditional distribution of precipitation (given precipitation occurs) (CAVG) and the coefficient of variation of the conditional distribution (CCV) can be used to estimate parameter values of this parameterization.

Historical values of CCV for observed precipitation tend to be near or slightly more than 1.0. Accordingly, a Weibull distribution usually fits the observed precipitation data fairly well. The distribution of precipitation forecasts depends on the skill of the forecast. If the skill is very high the forecast distribution may resemble the precipitation distribution. As forecast skill diminishes, the mean of the forecast approaches the climatological mean, the CCV diminishes and the POP increases. We generally use either a Gamma or Weibull distribution to parameterize the forecast climatology. Note that when CCV = 1.0, the Weibull, Gamma and Exponential distributions are identical. So if CCV is near 1.0 the exact choice of distribution is not a major issue.

The seventh parameter of the joint distribution model is a correlation parameter. The climatological distributions are used by the model to map observed and forecast values into a special "anomaly" space where the climatology of the anomalies are standard normal deviates. This anomaly space is illustrated in Figure 2 where variables u and v are standard normal deviate transformations of forecasts and observations, respectively. The curves shown in Figure 2 are contours of the bivariate standard normal joint density function f(u,v).

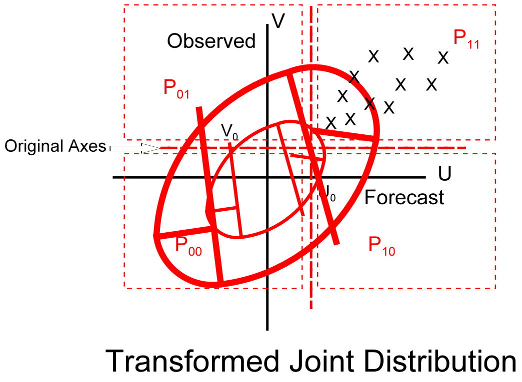

Figure 2 - Joint distribution transformed to standard normal deviate anomaly-space

Correlation parameter (RHO) is the correlation coefficient of this transformed joint distribution. This parameter is the only parameter that represents the skill (i.e. resolution) of the forecast. The other parameters control the reliability or calibration of the forecast.

The conditional distribution of future precipitation to be expected for a given forecast is derived by first transforming the forecast to a forecast anomaly (variable u in Figure 2) using the marginal forecast climatology distribution. Then a conditional distribution, f(v|u), for the future precipitation anomaly is derived from the bivariate standardized normal probability distribution using parameter RHO. Finally the conditional future precipitation anomaly distribution, f(v|u), is transformed to a conditional precipitation distribution using the marginal precipitation climatology distribution.

Note that this approach compensates for systematic differences between precipitation forecasts and precipitation observations.

3. NWS PARAMETER ESTIMATION STRATEGY

The parameters to be estimated from archived forecasts and observations are:

Forecast Climatology:
POP_{fcst}
$CAVG_{fcst}$
CCV_{fcst}

Observed Climatology
POP_{Obs}
$CAVG_{Obs}$
CCV_{Obs}

Correlation Parameter
RHO

These statistics vary spatially, seasonally and with forecast lead time. All of the statistics are scale dependent . An important application decision is the space and time scale at which this technique will be applied. These scales should be large enough to capture most of the skill in the forecast. One of the limitations of this technique is that forecast skill for aggregate precipitation amounts larger than the application scale is lost. Accordingly, ensemble forecasts will tend to over estimate uncertainty at larger scales. Conversely at smaller scales there is more uncertainty than would be included in a deterministic down scaling of the ensemble members. Accordingly, some kind of statistical down scaling may be useful to increase uncertainty at smaller scales.

Initial NWS application strategy is to apply the technique directly to each RFC sub-basin and to each 6-hour computational time step. The conditional precipitation forecasts are used to re-scale historical events independently for each sub-basin and each time step. But the internal pattern structure of the historical events assures that the Pearson rank correlation structure of all properties of the ensembles are exactly the same as in the historical data. This includes joint relationships between precipitation and temperature and between any arbitrary combination of points in space and time - anywhere in the U.S. But issues of scale dependency of the forecast skill need to be evaluated and better understood.

Since the length of historical archive is very limited (only a few years), it is essential to use "neighboring" values in time (and possibly space) to estimate model parameters. Smoothed climatological statistics for each RFC sub-area are estimated using a 45-day window both before and after a given day of the year. This assures smoothly varying parameter values seasonally during the year. Additional spatial smoothing of the local seasonally varying parameters may be desirable and is being considered.

Initial experience with the technique has been positive (Mullusky et al, J5.5 this Conference), but there is some evidence that parameter uncertainty may be large enough to limit the skill of the hydrologic forecasts. In one case the local parameters for one of the RFC sub-areas had to be replaced by the parameters for a neighboring sub-area.

One way to investigate the effect of limited archive data on parameter uncertainty and the subsequent effect of parameter uncertainty on forecast resolution (skill) and reliability is through numerical simulation experiments. That approach is used in this study as described below.

4. DESIGN OF A NUMERICAL SIMULATION SYSTEM

A system to conduct numerical simulation experiments has been developed and used to test the potential effect of limited historical archives of forecasts and observations on parameter uncertainty and on ensemble forecast verification statistics. The rationale behind this system is that if the statistical relationship between forecasts and observations is similar to that assumed by the NWS ESP short range preprocessor and ensemble generator, then it can be used to simulate the way data are archived and used to forecast future events. The numerical simulation process is as follows:

4.1 *Assume "True" Parameters*

Parameter uncertainty and degradation of ensemble forecast verification statistics resulting from limited length of historical archives depends on the climatology of events and on the skill of the deterministic meteorological forecast. For a given climatology and level of forecast skill the numerical simulation system can estimate the effect of length of historical data archive. Therefore, values of the "true" model parameters must be assumed and are provided as input to the system. Many different assumptions can be made and the system will provide results for each assumption. Below, results will be presented for one typical set of assumed model parameters.

These assumed parameters define the "true", but unknown, forecast environment.

4.2 *Generate Long Reference Set of Observations and Forecasts*

The joint distribution model described above is used to generate a very long set of pairs of forecasts and observations consistent with the model. This reference data set represents an unconstrained length archive that would never be practical to achieve but where there is very little effect of parameter uncertainty on the results. Statistics of the reference data are computed to assure that the data are consistent with the assumed model parameter values.

Each reference forecast is then used to generate a precipitation ensemble forecast. These ensemble forecasts are then verified using the observed reference values. The reference verification statistics represent the limit of how well the forecast system can be expected to perform for the given climate scenario and underlying skill (defined by RHO) of the deterministic meteorological forecasts. Therefore, the results of subsequent simulation experiments will be compared to the results of the reference data set.

4.3 *Simulate Replicate Archive Samples of Observations and Forecasts*

Different lengths of historical archive are considered. For each archive length, many replicate simulations are made. Parameter values for each replicate archive for each archive duration are computed.

4.4 *Estimate Parameter Values from Sample Archive and Simulate Future Forecast Operations*

Each set of simulated archive data are used to estimate parameter values and to simulate future forecast operations that could occur using each estimated parameter set. The forecast simulation involves generating an additional long period of simulated future deterministic forecast and observation pairs consistent with the "true " model parameter values. Then, estimated parameters from the simulated archive are used together with the additional long period of simulated future forecasts to generate precipitation ensemble forecasts. The ensemble forecasts are then verified using observations from the additional long period of simulated forecast and observation pairs that are consistent with the "true" forecast environment.

4.5 *Compute Summary Statistics for Each Archive Record Length*

Summary parameter and ensemble verification statistics are aggregated over the set of replicate archives for each archive record length. As the archive record length increases parameter uncertainty is expected to decrease, uncertainty in verification statistics should diminish and average values of verification statistics should improve.

5. ENSEMBLE FORECAST VERIFICATION STATISTICS

Several different ensemble verification statistics were used in this study to illustrate how parameter uncertainty may affect forecast resolution and reliability. This is not meant to be an exhaustive analysis of all possible verification statistics. The few used in this study are explained briefly below.

5.1 *Ensemble Mean Skill Score*

A statistic that is closely related to the correlation coefficient is the Nash-Sutcliffe efficiency statistic. In the verification literature this statistic would be called a skill score because the value of the statistic is scaled by the climatological variance of the observations (Wilks, 1995). The statistic is

$$\mathbf{EnsSS} = 1 - \frac{(1/n)\sum(\text{ENSavg-OBS})^2}{(1/n)\sum(\text{OBS-OBSavg})^2} \qquad (1)$$

where,

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

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

If the forecasts are unbiased (I e. forecast mean and standard deviation is the same as observed) 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.

This skill score can be applied to measure the skill relationship between the mean of the entire ensemble and the observed value. This measures the aggregate skill of the forecast. It also can be applied to the conditional part of the forecast that gives the conditional probability of precipitation if precipitation should occur. The conditional ensemble mean skill score **CEnsSS** measures the skill relationship between the mean of the conditional distribution of precipitation and the observed value, conditioned on events when precipitation occurs.

5.2 *Brier Skill Score*

A measure of accuracy of the forecast probability of precipitation is the Brier score (Wilks, 1995). The Brier skill score is defined as

$$\mathbf{BSS} = 1 - BS / BSC \qquad (5)$$

where BS is the Brier score,

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

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,

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

where p is the climatological probability of the event.

5.3 *Measures of Forecast Reliability*

The reliability of probability forecasts (conditioned on precipitation occurring) can be assessed by constructing what is known as a reliability diagram (Wilks,1995). 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 the reliability diagram and the observed relative frequency is taken from the forecast probability associated with the observation.

The magnitude of the vertical deviation of the observed relative frequency from the 45-degree diagonal is a measure of reliability (Wilks, 1995). In this study we using the RMS value of this deviation as a measure of reliability of the forecast probabilities,

$$\textbf{RMScens} = [(1/n) \sum (\text{Fobs - Funiform})^2]^{1/2} \qquad (3)$$

where n is the number of ensemble forecasts corresponding to observed precipitation events.

A similar reliability diagram can be constructed to measure the reliability of probability of precipitation (POP) forecasts. This involves use of several POP forecast categories. The corresponding statistic is **RMSpop** that measures the RMS deviation of observed relative frequency of precipitation occurring from the mean POP forecast value for each category.

6. EXAMPLE SIMULATION RESULTS

An example numerical simulation was made for a given climatological scenario and for a relatively highly skillful short term precipitation forecast. The assumed parameters are in Table 1.

These parameters are typical of daily precipitation. The daily mean precipitation would be 0.20 * 10 = 2 mm/day or 730 mm/yr. In this example, the forecast and observed climatologies are assumed to be the same, but they could have been different without major effect on the results. The anomaly correlation between forecasts and observations is assumed to be 0.80. The effect of parameter uncertainty for less skillful forecasts has not been studied.

Table 1 - Assumed Parameter Values

Parameter	**Value**
POPobs	0.20
CAVGobs	10.0
CCVobs	1.0
POPfcst	0.20
CAVGfcst	10.0
CCVfcst	1.0
RHO	0.8

Different lengths of forecast archive were considered in multiples of 30-day (i.e. monthly) increments. The multiples were taken as the geometric series of 1,4,16,64 and 256 months. It is assumed that there are 30 independent events in each month. The number of replicates of parameter values for each archive duration was decreased with increasing archive duration because uncertainty diminishes with increasing archive length. Therefore, fewer replicates are needed to achieve the same accuracy in the simulation result. The number of replicates followed the reverse geometric series with 256 replicates for a 1-month archive and only 1 replicate for the 256 month archive.

The reference data set also had a duration of 256 months. All of the results from the simulated 256 day archive were very close to the results for the reference data set so it is clear that the uncertainty in the longest simulated archive is negligible.

The main interest in this study is in how parameter uncertainty caused by limited length of forecast data archives affects forecast performance. One illustration of the effect of limited archive length on model parameter is shown in Figure 3. The standard deviation of error in parameter **CAVGfcst**, expressed as percent of the mean, decreases from 45 percent for a 1-month equivalent length archive to less than 10 percent for an archive of 12-months or more.

Parameter uncertainty has two effects on ensemble verification statistics. One is to diminish the mean value of skill scores. The other is to introduce uncertainty in skill scores so there is a substantial chance a given archive could reduce the skill score by chance owing to the shortness of the archive. It is possible to estimate both effects by making replicate simulations for each archive duration.

Figure 4 shows the “reliable” value of **EnsSS** vs archive length. The “reliable” skill score shown in Figure 4 is equal to the mean skill score minus one standard deviation of the skill score uncertainty. If **EnsSS** is negative it would be better to replace the entire precipitation ensemble forecast with the climatological distribution than to use the ensemble derived from the atmospheric forecast

Figure 5 shows how the “reliable” value of the skill score for predicting the mean value of the conditional part of distribution of precipitation, **CEnsSS**, depends on archive length. This figure shows that skill in short range ensemble precipitation forecasts of probability of precipitation amounts can be very significantly reduced unless there is adequate archive data to estimate model parameters. If **CEnsSS** is negative, it would be better to replace the conditional part of the ensemble forecast with the climatological distribution of wet precipitation events.

Figure 6 shows how the “reliable” value of the Bier skill score (**BSS**) depends on the archive record length. This is the average **BSS** minus one standard deviation of the uncertainty in the **BSS**.

Figure 7 shows how the reliability of conditional precipitation probability forecasts **(RMScens)** varies with archive length. The probability RMS error in this figure is equal to the mean error plus one standard deviation of the uncertainty in the error.

Figure 8 shows how the reliability of probability of precipitation forecasts **RMSpop** depends on archive length. The RMS values shown are the mean error plus one standard deviation of the uncertainty in the error

Together, Figures 3-8 suggest that at least a one year-equivalent length of independent daily observations is needed to reduce the effect of parameter uncertainty to a minimum level. This result depends on the assumed climatology and on deterministic forecast skill. Because climatology and forecast skill vary seasonally, it is not possible to get 365 independent daily observations from a single calendar year archive. At best it may be possible to combine days together in a 90 day window to reduce the required actual archive length to about 4 calendar years.

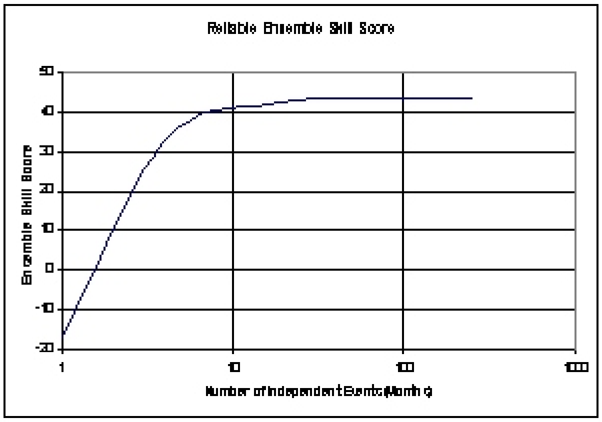

Figure 4 - Effect of archive length on ensemble skill score **(EnsSS)**

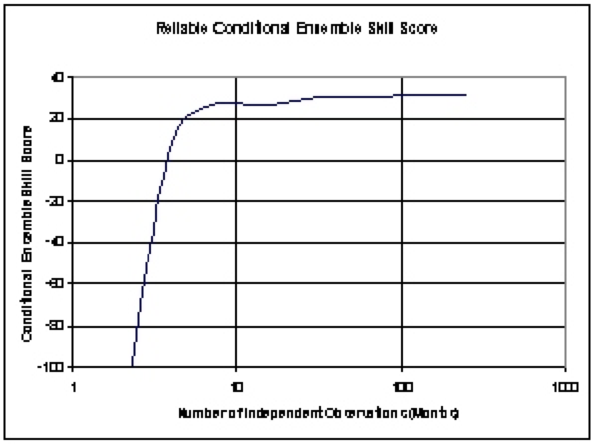

Figure 5 - Effect of archive length on conditional ensemble skill score **(CEnsSS)**

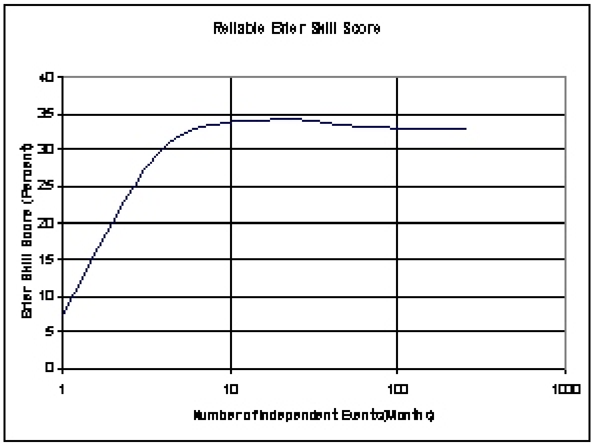

Figure 6 - Effect of archive length on Brier skill score **(BSS)**

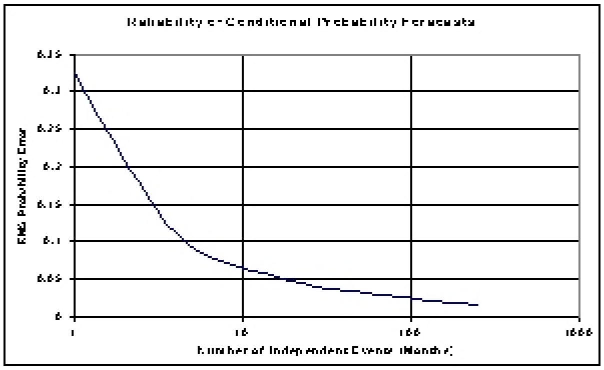

Figure 7 - Effect of archive length on reliability of conditional precipitation probabilities **(RMScens)**

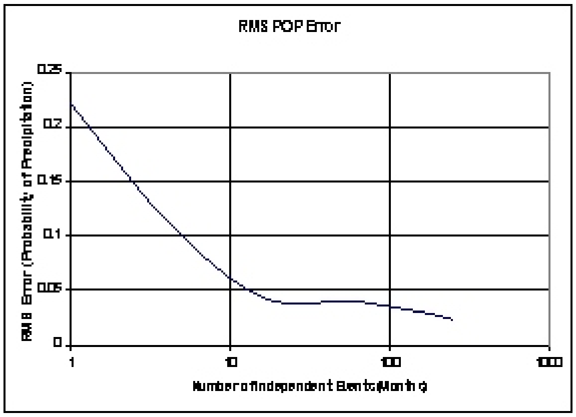

Figure 8 - Effect of archive length on reliability of probability of precipitation forecasts **(RMSpop)**

7. CONCLUSIONS

Reliable ensemble streamflow prediction requires unbiased ensemble precipitation forecasts as input to hydrologic forecast models. Statistical preprocessing techniques can remove biases in meteorological forecasts to meet this requirement. But these techniques have parameters that must be estimated using historical forecast archives. The resulting preprocessing techniques are only valid if they are applied to forecasts from the same system that produced the archive.

A numerical simulation procedure can be used to assess how limited archive length influences both parameter uncertainty and skill of ensemble precipitation forecasts derived from deterministic meteorological forecasts.

About 4 calendar years of forecast archive are required to minimize the effect of uncertainty in parameter values on forecast verification statistics. Skill in forecasts of conditional probability of precipitation (**CEnsSS)** are very sensitive to short lengths of forecast archive. These results depend on the assumed climatology and deterministic forecast skill. Additional study is needed to understand how the required archive length depends on these assumptions.

6. REFERENCES

Mullusky, M., L. Wu, X. Fan, J. Ostrowsky, N. Pryor and J. Schaake, 2003, Short term ensemble river forecasts: Application, AMS 17th Conference on Hydrology, Long Beach, Paper J5.5.

Wilks, Daniel S., 1995, Statistical methods in the atmospheric sciences. Academic Press, 467pp.