

# FINAL REPORT

---

## REAL-TIME VERIFICATION OF SHORT TERM ENSEMBLE HYDROLOGIC FORECASTS

PI: BISHER IMAM

CONSULTANT: HOLLY HARTMANN

NWS-TECHNICAL LEAD: KEVIN WERNER

STUDENT: ERIN PRITCHER

*Center for Hydrometeorology and Remote Sensing (CHRS)  
Civil and Environmental Engineering  
The University of California, Irvine  
Irvine, CA*

To verify: to establish the truth, accuracy, or reality of .... (M.W dictionary)

---

## I. BACKGROUND

---

### I.1 FORECAST VERIFICATION

Forecast verification is a process that aims to quantify and summarize the relationship between forecasts and observations. It also includes the problems of comparing observations, forecasts, and a reference forecast, particularly when an attempt is being made to compare different forecasts and/or forecasting systems (Potts, 2003). Wilks (1995) defines forecast verification as the process of determining the quality of forecasts. This requires the utilization of quality measures that summarize one or more aspects of the relationship between forecasts and observations. Technically, the three main objectives of forecast verification are (a) monitoring quality, (b) improve quality, and (c) compare the quality of different forecasting systems. However, users of forecast verification results range from administrators, who want to know the value of investing in forecast system improvement to forecasters and modelers, who want to assess areas of improving their own predictions, to forecast users, who weigh their decision based not only on the forecast but also on the quality of such forecast.

By definition, the verification problem is a posteriori problem in the sense that it requires the simultaneous availability of observations along with their matching forecasts. In other words, one can not literally verify forecasts in realtime because their pertinent observations remain in the future. Therefore, realtime verification of forecasts must then be defined in terms of “quality of the forecasting system” as opposed to the problem of measuring the quality of the specific forecast. By referencing the forecasting system to specific conditions, the realtime verification problem becomes an assessment of the system’s performance as demonstrated by previous forecasts that were issued under conditions similar to those of the period immediately preceding the current forecast (i.e, conditional verification).

Since the Finley affair (1884-1893), which is considered as the starting point of developing verification measures and methods (Murphy, 1996), much of the developments in verification theory in the Earth sciences continued to occur almost exclusively within the weather forecasting discipline. Hydrologists, on the other hand, have for long time used various forecast quality measures for model calibration and validation studies (Viessman et. al., 1970). There has been considerable interest in verifying deterministic forecasts over the years. The current hydrologic verification system, although very rudimentary, it based strictly on deterministic forecasts. As such, both forecasters and administrators are familiar with many verification measures, particularly those associated with deterministic forecasts. Efforts to streamline, improve and re-organize hydrologic verification have with the establishment of a NWS verification team. The team’s report includes a review of both deterministic and probabilistic forecast measures, and recommendations for inclusion of a selected subset of measures into the next-generation verification team. Clearly, the notion of verifying probabilistic hydrologic forecasts has taken hold after the development and adoption of probabilistic forecasting approaches. The increasing utilization and popularity of probabilistic hydrologic forecasts, which were pioneered by the National Weather Service in

the form of Ensemble Streamflow Predictions (ESP) (Day, 1985), have resulted in various studies aiming to incorporate verification into hydrologic practices.

Our discussions with several forecasters and hydrologists in charge at various RFCs indicated that operational hydrologists view verification in a broader sense than their counterparts within the meteorological community. Their view encompasses the utility of verification in validating whether a forecast is ready to issue as a final product or determining that it needs more work. In this view, verification becomes an operational issue as opposed to purely diagnostic issue. Such poses significant challenges including the determination of sample size, the identification of conditional verification (i.e., similar conditions), tracking model states, input, and output, relative to their climatology, and the establishment of links between the forecast issuance, verification, and simulation component of the forecast system. The above mentioned observation also highlights the necessity of conducting verification in the manner of which forecasts were issued. In this regard, the availability of archives of actual forecast is the best guarantee of appropriate verification study. In case such archive is not available, and one must resort strictly to performing re-forecast simulation (i.e. hindcasts), it is important that hindcasts whether deterministic or probabilistic, attempt to simulate, the largest possible extent, the actual forecasting procedure. This requires simulating forecasters' judgments (e.g., modification to estimate most appropriate initial conditions, quality control), or a best approximation of the impact of her/his judgment on the forecast. This begs for distinguishing between research aiming at furthering verification theory, and that aiming at improving the utility of verification in operations. In the latter, researchers must be knowledgeable with the standard operational procedures, and able to use the same system the forecasters use. This has been a guiding principle of this study. We have attempted to the largest possible extent, to utilize the NWSRFS, its tools, and to consult with NWS staff and forecasters. For example, consider figure 1, which summarizes what we understand to be the major steps taken by a forecaster in the CNRFC to perform the major forecasting task for the day using NWSRFS. Clearly, identification of initial appropriate initial condition is the primary task associated with both deterministic and probabilistic forecasts. On the other hand, probabilistic forecasts (ESP) are automated without forecaster's interference. Yet, they provide more accurate link to the likely model states on that day than those provided by "historical simulations".

Verification is fundamentally a statistical problem with roots in regression, probability distribution and multi-variate analysis theories. As mentioned earlier, most verification methods involve calculating a measure or a suite of measures that summarize various characteristics of the relationship between predictions and observations. In addition to numerical measures, exploratory (visual) analysis also serves as a complementary tool. In deciding which verification measure should be utilized, one must account for factors such as the type of forecasts, the predicted event, and the relevant forecast quality attributes. Figure 1, summarizes the key elements of forecast verification problem.

This report includes, in addition to the contractual deliverables, a review of the NWS-Verification team draft report, along with two surveys intended for distribution to NWS forecasters and HICs. These additional deliverable were developed in collaboration with our project's technical lead in order to facilitate future design and implementation of verification system. The surveys were provided to Union representatives and their comments were fully incorporated.

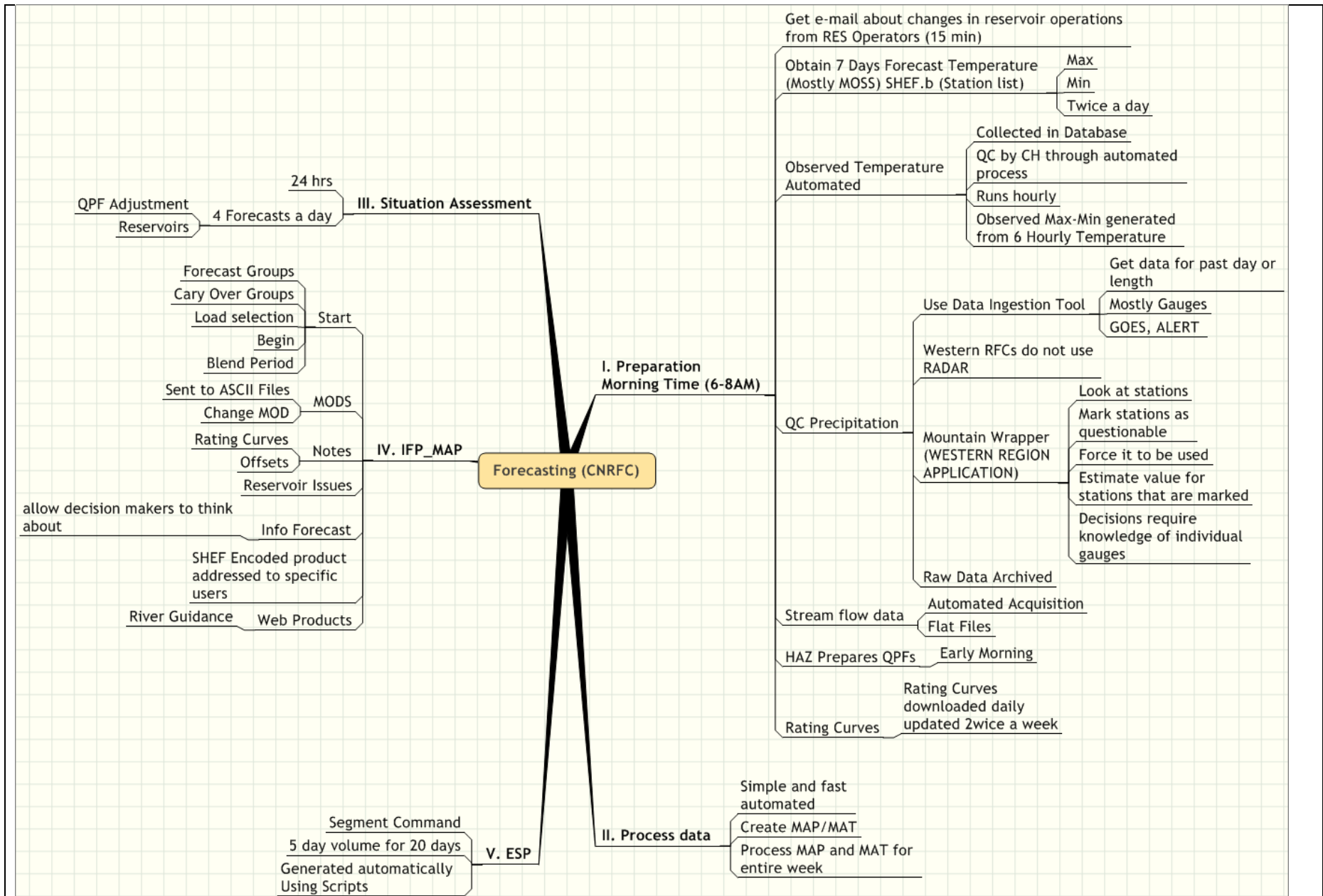


Figure 1. Forecasting procedure used by a representative CNRFC forecaster. Notice the amount of efforts that is paid to the development of appropriate model initialization. Notice also that ESP production, in this case is influenced by the deterministic forecast initial condition, but is automated.

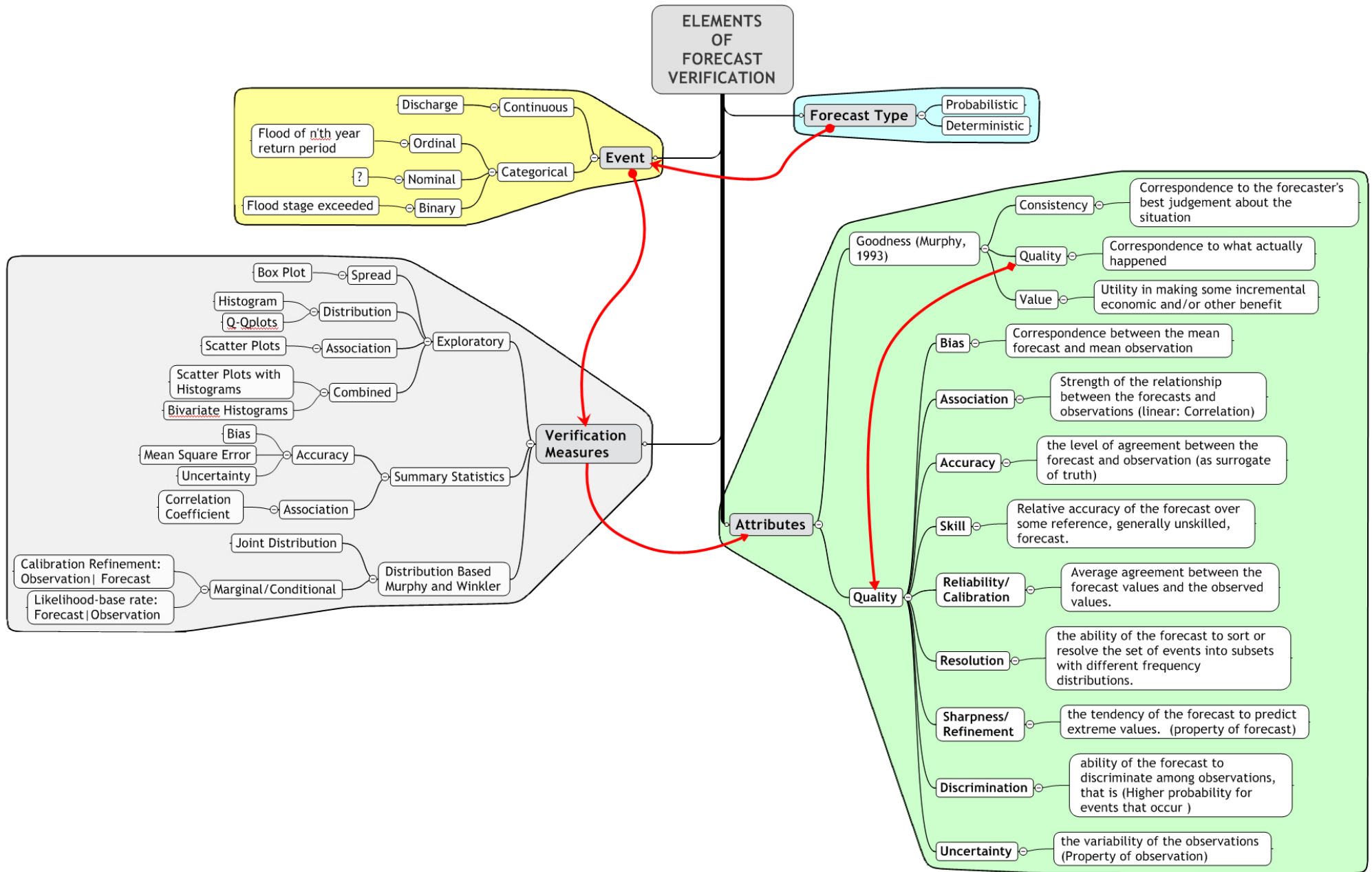


Figure 2. Elements of Forecast verification problem (Predicted event, type of prediction, forecast attributes, verification measures).

---

## II NWSRFS-ESP

---

### II.1 OVERVIEW

This project focuses on short term (up to 15 days) probabilistic streamflow forecasts, particularly those generated using the National Weather Service River Forecast System's Ensemble Streamflow Predictions (NWSRFS-ESP) procedure (Figure 3). The traditional implementation of ESP is based on using historical traces of precipitation and temperature, along with other relevant information (e.g. reservoir releases). This historical data are used to force the hydrologic model with initial conditions identified by the most recent observations of forcing variables. The result is an ensemble of traces representing possible discharges conditioned by the current initial conditions. In some river forecast centers, ESP traces are forced with probabilistic quantitative precipitation and temperature forecasts, particularly for the first few days of the forecast. However, this approach is still being developed and has yet to be implemented system wide. NWS is interested in issuing and utilizing short term probabilistic forecasts, where the duration of the forecast is less than 15 days. These forecasts can take a variety of forms, e.g., exceedence/non-exceedence probabilities of total volume for a given duration, or categorical forecasts such as the probability of exceeding certain flood stages.

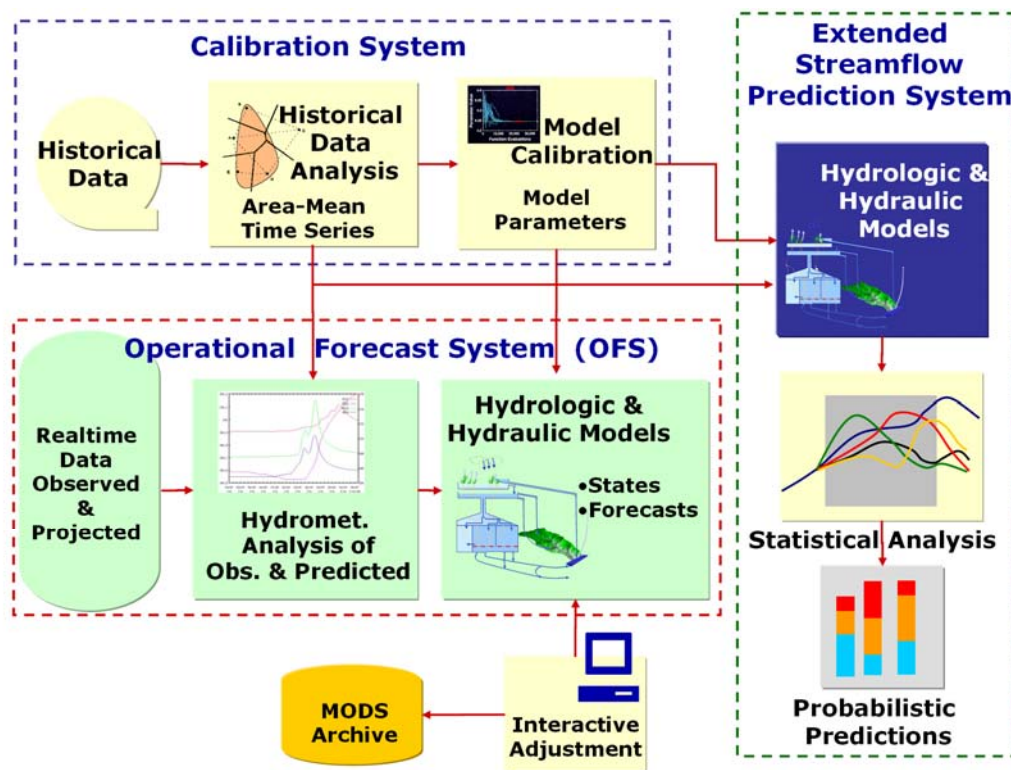


Figure 3. Schematics of the NWSRFS including ESP component.. Again notice the sequential independence of ESP forecast creation process from the deterministic forecast.

In order to obtain statistically significant verification results, particularly for ensemble forecasts, a large number of forecast ensemble and corresponding observations are necessary. These can be obtained by hindcasting (e.g. re-forecast), which is made possible through one or more of the tools available in the NWSRS (i.e., etsgen tool). The following discussion presents the approach used to generate the re-forecast information presented in this report.

## II.2 CASE STUDY

We selected the Minnesota Forecast Group within the North Central River Forecast Center NCRFC for this study (Figure 4). The selection of this particular sub-basin was agreed upon during a meeting with NCRFC. As seen in the figure several headwater basins are available for the study. The selection of headwater basins aims at *dealing with forecast verification at the forecasting unit level*. This is particularly important because modifications, which are known as “MODS”, are generally carried out at the forecasting unit level.

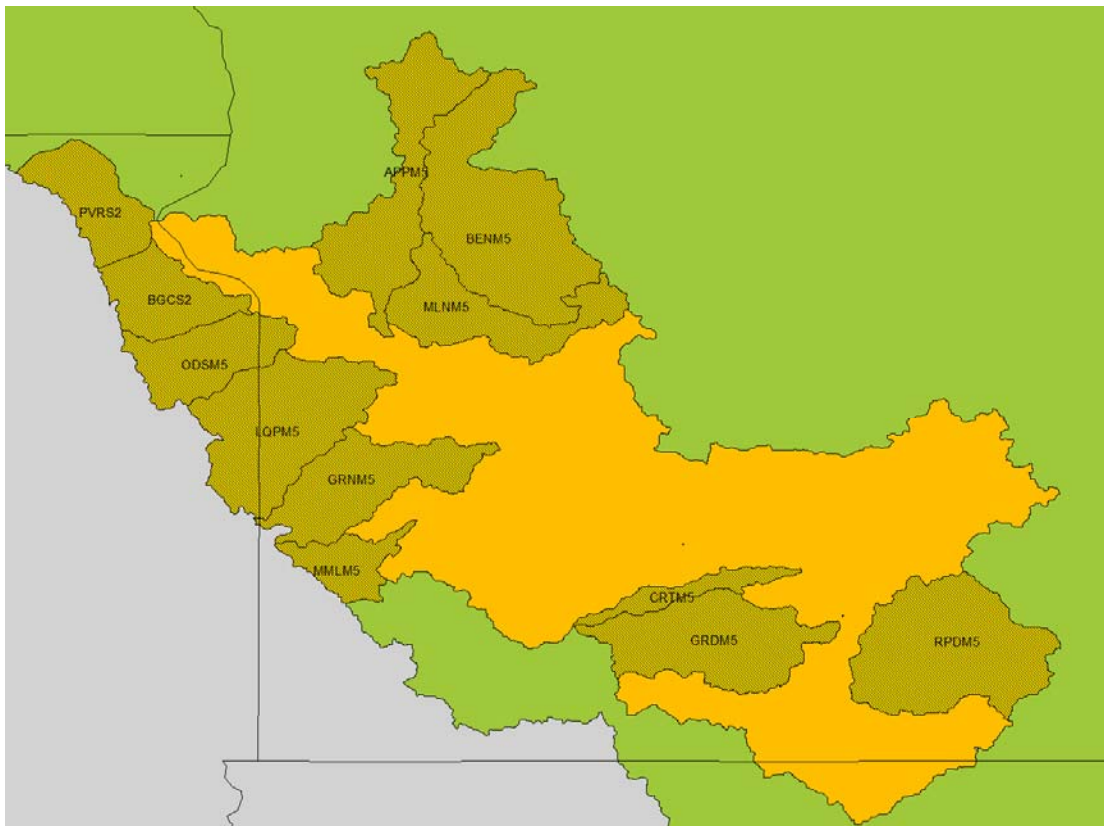


Figure 4. NCRFC study area within the Minnesota River Basin.

The key criterion affecting the selection of a given basin for diagnostic verification study is the availability of long-term precipitation (MAP), temperature (MAT), 6 hourly instantaneous discharge (QINE), and daily average discharge (QME). Therefore, our first task was to explore the available data for each of these variables.



Because our task aims at verification against actual historical data as opposed to historical simulation (known as HS files), two considerations were important. First, the historical records, which are available in the /calb/ (calibration data) directory must constitute both the ensemble forcing (MAP, and MAT) and the observations required for the creation of a verification data set (QME and QINE). Considering that our focus is on short term predictions, verification of probabilistic forecasts of cumulative volumes will be less important than verification of daily discharge. Figure 5, illustrates the average daily discharge data available for the same set of basins. Our selection included:

- *Lac Qui Parle River [LQPM5] near Lac Qui Parle MN (USGS station 05300000)*
- *White Stone River near Big Stone City-SD, [BGCS2] (USGS station 05291000)*
- *Pomme De Terre River at Appleton MN [APPM5] (USGS Station 05293000)*
- *Le Sueur River Near Rapidan MN [RPDM5] (USGS station 05320500).*

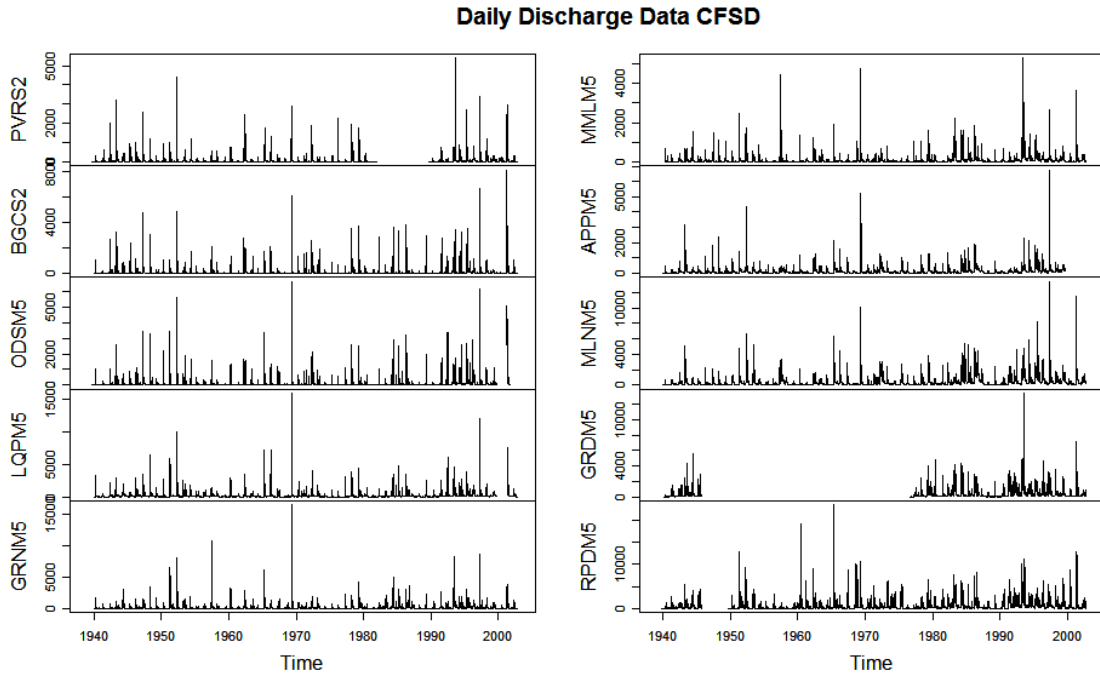


Figure 5, average daily discharge data for the study area

The next step is to determine the study time-frame. To do so, we considered the statistical properties of the historical timeseries. Figure 6, shows a box-whisker plot for the LQPM5 watershed mean daily flow based on day of year with very large outliers being truncated above 4000 CFSD. Clearly, the most critical period for real-time short term forecast that is consistent with flood forecasting is the time period between March 1<sup>st</sup> and July 15. Similar hydrographs are seen for most of the other headwater basins. This is consistent with snow-melt season and early summer convective storms. Therefore, we proceeded to create a diagnostic verification data set forecasts verification as seen in figure 6. This includes defining a forecast window (30 days) to allow for monthly flow volume verification if needed, a forecast stagger (1 day) to allow for future consideration of lead time impact on verification measures, and forecast analysis (15 days) to capture short term



forecasts. Carryover times will be saved on the each day of every month during the time frame and ensemble streamflow forecasts were generated henceforth. Ensemble simulations for the entire period have been completed for RPD5, with other basins in progress for future studies.

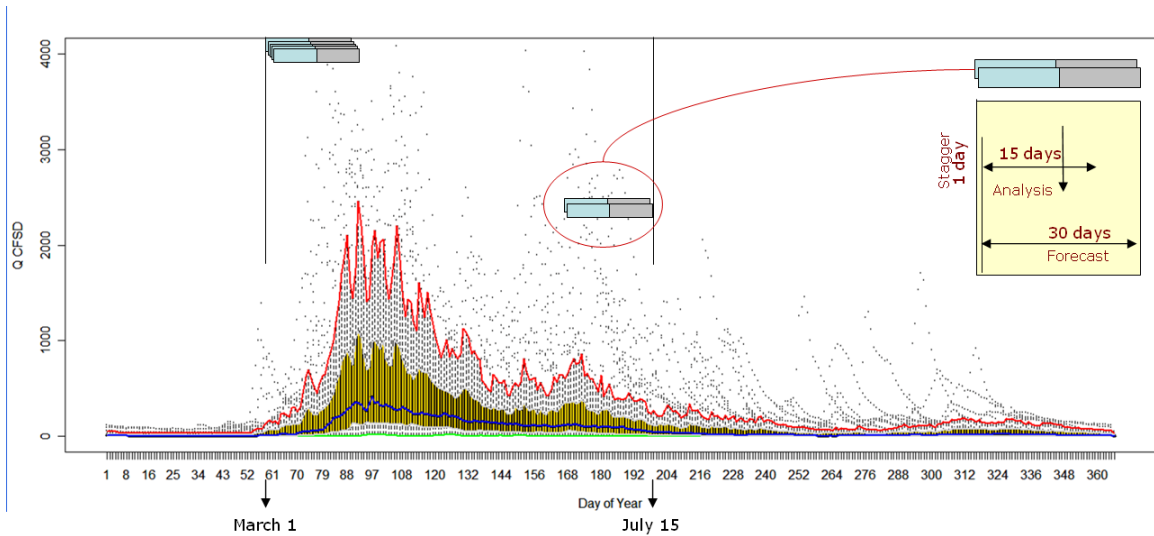


Figure 6. Experimental design for the report's diagnostic forecast verification study. The box on the right shows the forecast, analysis, and stagger periods. The runs were performed to accommodate future expansion of research objectives including lead time impacts.

The system installed at CHRS had several issues. Among these issues is the lack of connectivity of some key java libraries required to complete several scripting tasks aiming at converting the data from ESP card format into formats that could be read by standard software packages (ASCII tables). In order to complete our analyses, we had to resort to alternative methods that utilize several other tools in order to complete this project. It must be mentioned that all hindcasts performed were generated automatically and have not been cross validated. In addition, few starting carry-over dates were available and the system was not connected to new data, precluding consideration of additional information generally used such as NEXRAD precipitation, and/or QPPF. We attempted to emulate, to the best possible extent, the operational procedure of the NWS. Alas, the lack of data update forced our analysis to utilize only the most elementary approach to generate ESP hindcasts, and therefore the results presented in this study are not diagnostic, nor are they to be considered a conclusive assessment of ESP. Figure 7 shows the configuration of the analysis system we implemented in our study.

As seen in the figure our analysis utilizes the historical data available in the /calb/ directory, particularly the area\_ts files. The analysis proceeds as following:

1. Using RTP's timeseries tool (*tstool*) convert the NWS Card files (QME and QIN) into ascii tables (comma or tab delimited). These files are then introduced into the R statistical package to develop the historical climatology data set for each day of the year (Figure 7).
2. Using etsgen, generate the ensemble forecasts for each carryover date (1 day stagger and 30 days long forecasts as mentioned in the previous section)

3. Again using RTI's *tstool* convert each BASIN.BASIN.SQME.24. YEAR.CODATE file into a corresponding ascii table (tab or comma delimited).
4. Using the R statistical package along with the results of step 3 extract verification data sets (as described by Bradley et al., 2003, and Bradley et. al, 2004). This involves converting the observations into a binary event using the climatological CDF, and identifying the probability forecasts for the selected day during forecasting period. Also compute the mean standard deviation of the probabilistic forecasts associated with each climatological threshold, which will be used to compute the continuous rank probability scores.
5. Using the NCAR's R' verification package perform verification analysis for each day's verification data set (given probabilistic forecast and binary observations). Plot attribute diagrams and other verification measures. Use the verification objects to identify other potential plots. Additional packages are also used such Hmisc to estimate empirical distributions, gplots and DAAG to control graphic devices and to generate screen shot-like R output.

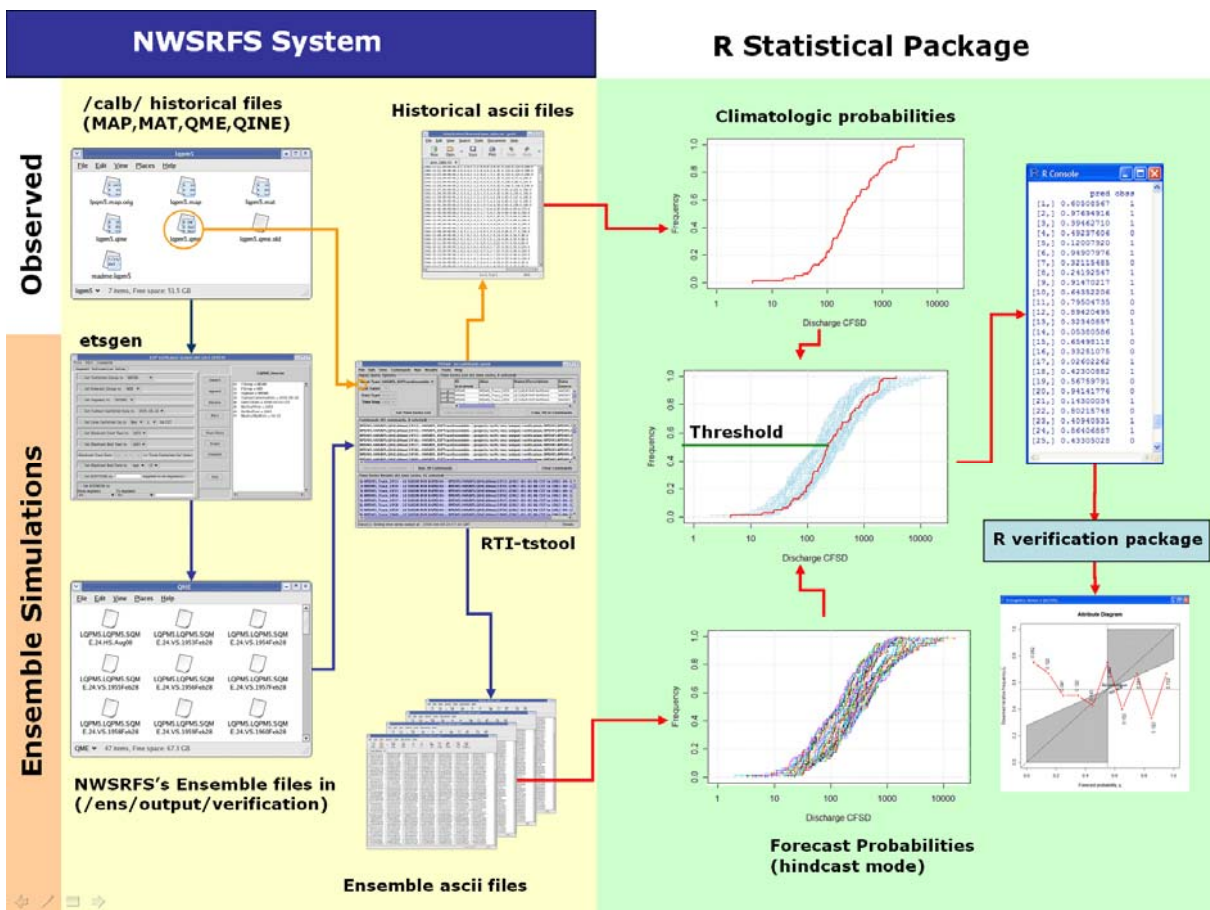


Figure 7, schematic diagram of analysis procedure. Reliance on manual procedures can be eliminated by considering a single RFC system instead of development version.

---

### III. EXPLORATORY APPROACH

---

#### III.1 INDIVIDUAL FORECASTS

Forecasters' ability to graphically explore recent observation and simulation data, recently issued forecasts, or a suite of hindcasts generated for verification purposes is essential in forming their initial understanding of the forecasting system. Each forecaster has a high level of familiarity with her/his forecast points, along with expertise in the capability of the deterministic components of the forecasting system. In the following, we first address these graphical approaches from the point of view of individual forecasts. It must be emphasized that while the examples presented herein use previous years of hindcasts, which were generated to exemplify quantitative verification measure and scores, the verification hindcast ensembles could easily be replaced with recent forecasts, or with specific forecasts associated with conditions similar to those existing at the time of forecast issuance. Needless to say, one must emphasize the exploratory nature of these graphical representations. Therefore, exploratory approaches can be utilized to address individual forecasts (ensembles) as well as a group of ensembles that are treated as a single sample.

##### 1. TRACE PLOTS

The first and most straight forward of exploratory graphics is the classical ensemble plot superimposed with observations. When these plots are available for each hindcast/previous forecast year of the same forecasting window, the forecaster can form an initial assessment of the "forecast-observations" pairs for each year and of the forecasting system's performance over the years. For example, consider Figure 7, which represent 15 hindcasts, each with 46 ensemble members, at the forecasting point (USGS 05320500 LE SUEUR RIVER NEAR RAPIDAN, MN, RPDM5: basin are 1110 square miles) during the analysis period April 1-15. This experiment, which will be used for the remainder of this report, was conducted by considering 45 year hindcasts, hereafter referred to as ensembles (1953-1997) with 46 ensemble members (hereafter referred to as traces) in each ensemble (1952-1997). From figure 8, the first benefit of side-to-side comparison of ensembles, along with observations (red line with dots) and the historical simulation (blue line with dots) is evident. A visual, qualitative assessment of the possibility of invalid probabilistic forecasts is possible, together with information about the magnitude and error associated with the model and/or initial condition using historical simulation. Trace plots could also be employed by forecasters to investigate the performance of recent forecasts. For example, during active precipitation/snow-melt period, forecasters could review the performance of recent short term forecasts, or the performance of the first few days of last week forecasts. In operational settings, this can provide the forecasters with a reasonable idea about the historical time-series ability to represent recent conditions, and to weigh the need for other sources of information. Notice that in figure 8, the presence of a distinct historical simulation trace (blue line with dots) which allows the forecaster to assess the magnitude of error associated with model initialization based on the historical simulation. For an experienced forecaster, such information is valuable. Furthermore, the figure highlights the need for an alternative ESP hindcast initialization strategy. This will be further discussed in later sections.

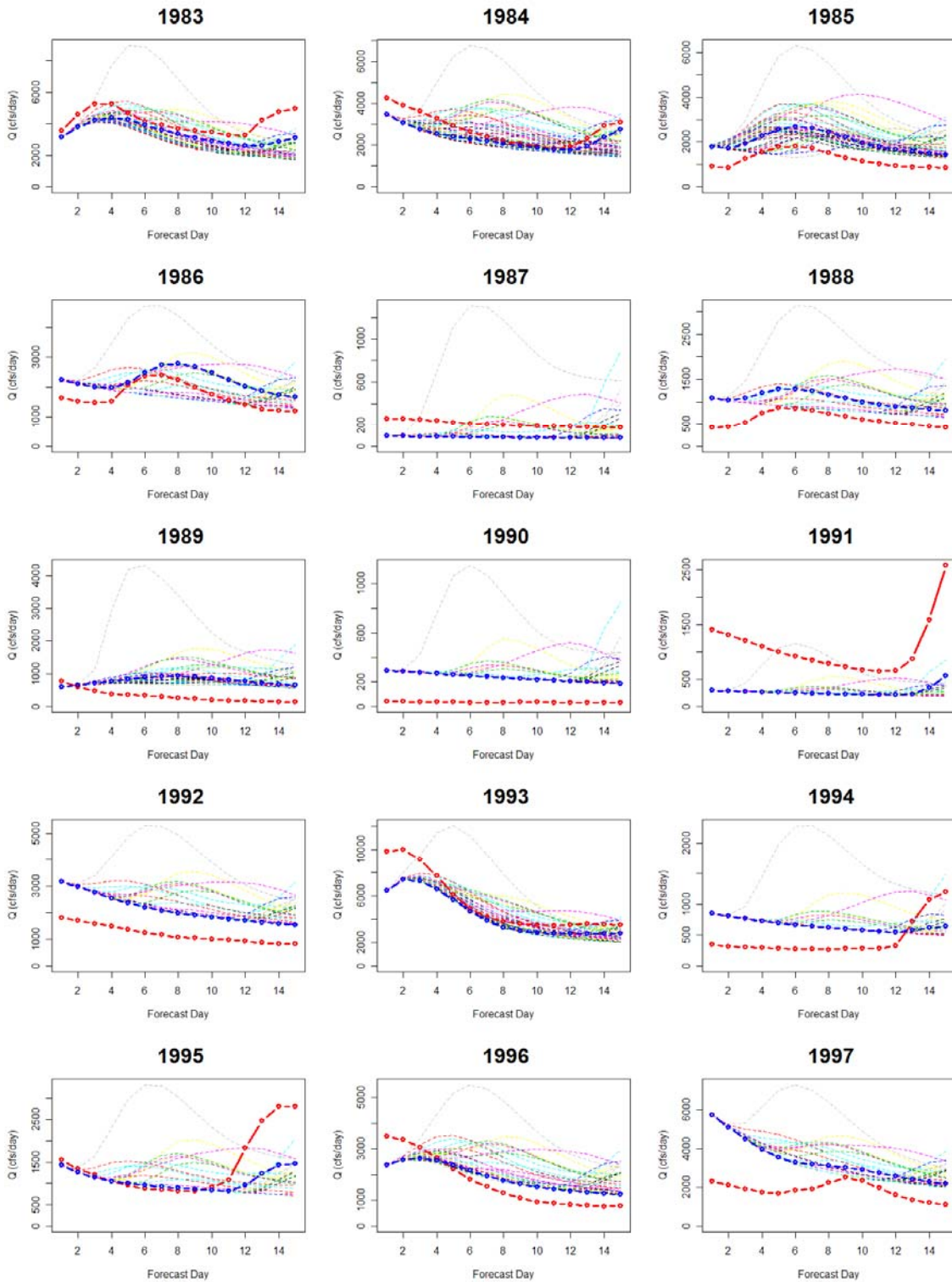


Figure 8. Sample screen shot of side comparison of trace-observation plots. Red and blue lines with dots: observations and historical simulation other lines: Traces of analysis period associated with the above-mentioned re-forecasting experiment.

## SCREENSHOT 1

Figure 8 also represents a possible screen shot that allows the forecaster to explore both the collection of hindcast ensemble, recent forecasts, as well as conditional verification data set. The forecaster can inspect the effects of initial condition on the disparity between the ensemble and the observations. She/he could also investigate the stability of recent forecasts of the same day but with different lead time, and finally to explore the possible selection of specific years within the conditional verification suite of ensembles. Clearly this screen of simple visual exploration of the data must be available to forecasters at any point during the verification. If the individual graphs are mouse-sensitive, forecasters can select the appropriate sub-set to include into more formal verification. The inclusion of both historical simulation, which is, in fact, a member of the ensemble associated with each year, allows the forecaster to obtain information regarding the potential source and magnitude of model error that may contribute to verification measures.

## 2. BOX-WHISKER PLOTS

While trace-observations plots provide qualitative visual representation of the ensemble's spread along with the observation, box and whisker plots can provide more quantitative visualization of the probability distribution of the ensembles, and therefore of the probabilistic forecasts themselves. The idea is not that the observations should fall within the center quartiles. Rather, the observations should occur throughout the ensemble distributions, at a rate that reflects perfect reliability. These plots are different than the standard (probability range-color) plots utilized by the NWS in the sense that no specific assumption regarding the probability distribution of the forecasts is needed. They represent sample statistics and quartile ranges. Similar to trace plots, B-W plots can also be employed by forecasters to investigate the performance of recent forecasts. Figure 9 shows box plots for the same data set used in Figure 8.

## SCREENSHOT 2

Figure 9 also represents a possible screen shot that allows the forecaster to explore both the collection of hindcast ensemble, recent forecasts, as well as conditional verification data set. In addition to inspecting the stability of recent forecasts, the forecaster could identify more qualitatively whether the observations fell within the range defining the probabilistic forecasts. Care must be exercised in this case not to expect observation to fall at the mid-range of the probability distribution of the forecasts. In fact, for forecast to have reasonable resolution, the observations should cover, collectively, various points within the probability distribution of the forecasts as summarized by the box-whisker plot.



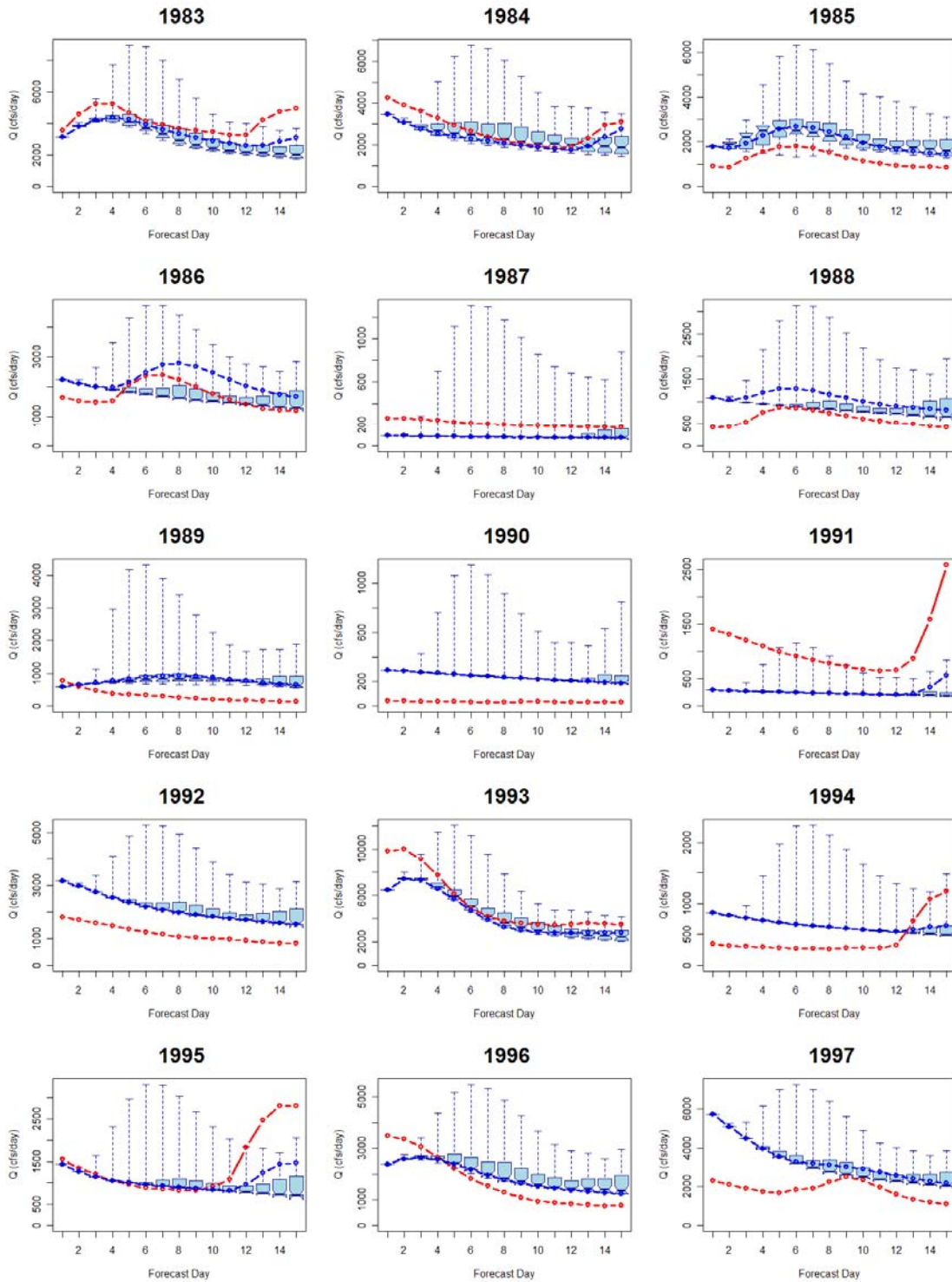


Figure 8. Box and Whisker plots for individual ensembles for the same data used in figure 9. Similarly, red and blue lines with dots represent observation and historical simulation, respectively.

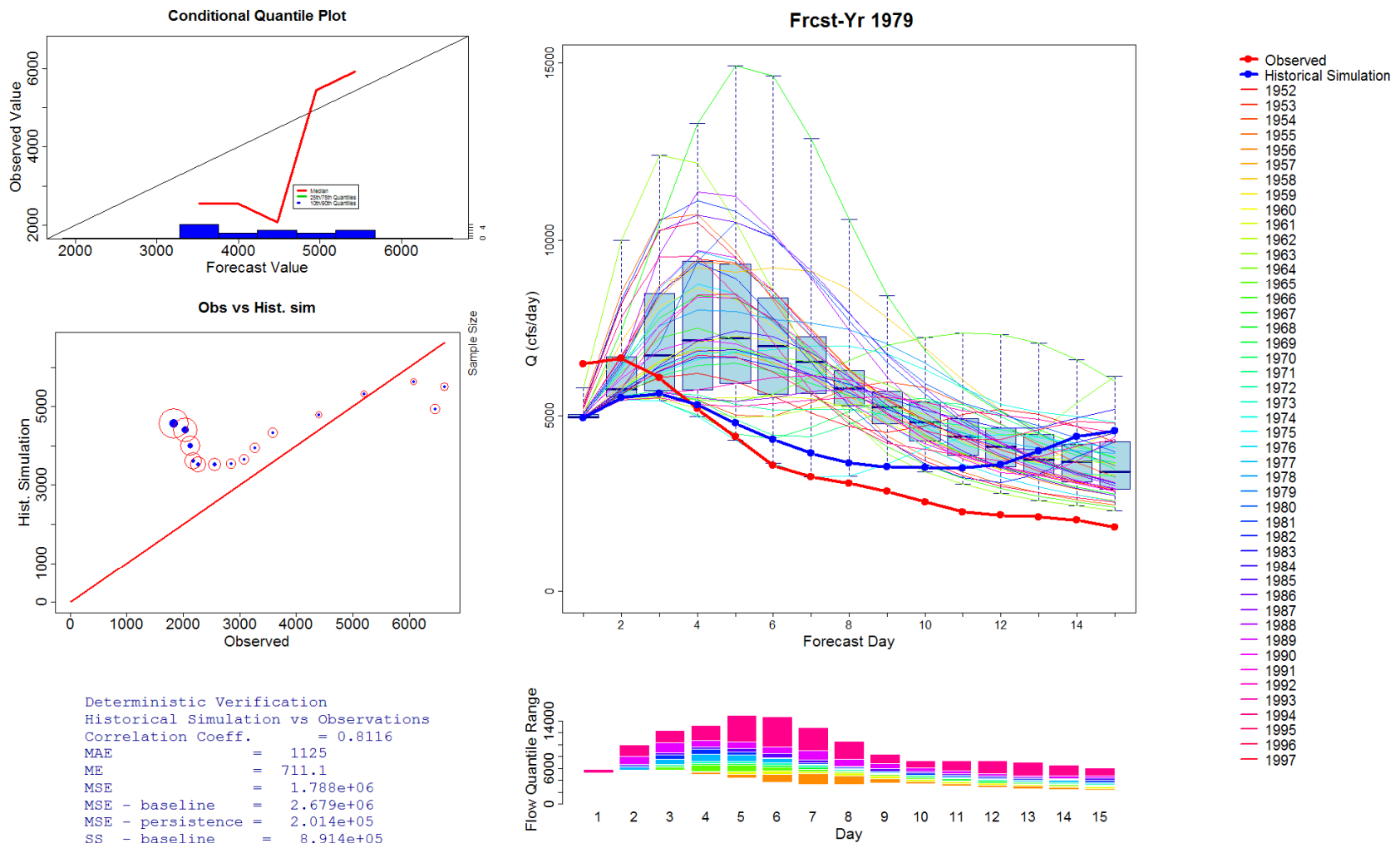


Figure 10. Combining box plots, trace-plot, observation and historical simulation can yield a reasonable amount of information to forecasters regarding a given forecast. Example shown is for year 1961. Notice that while the historical simulation nearly matches the shape of the observed data, significant error in the initial state of the flow is seen. Notice the presence of information regarding the deterministic verification of a single trace, which is the historical simulation trace.



### SCREENSHOT 3

Figure 10 shows a possible “detailed” exploration of a given forecast. In the figure, box-plot and trace plot are superimposed with the observation and historical simulation, but for individual forecast (top center). In addition, the probabilistic forecast is shown, and the forecaster can see whether the observations ensemble is providing sufficient probability bins (same as in espadp tool). In this case, 10 equally spaced bins are used. The rightmost column has the index color for each ensemble member year. The left column is the most important part of this screenshot. At the top, a deterministic verification of the historical simulation against the observation provides invaluable information to forecasters. As seen in the figure, this information includes the quantile-quantile plot (top-left panel), along with the observed-predicted scatter plot, with the size of the dots being proportional to absolute error associated with each pair of observed-historical simulation points (center-left panel). The lowest-left panel provides a set of deterministic verification measures of the historical simulation. These include Mean Absolute Error (MAE), Mean Error (ME), Mean Square Error (MSE), and Root Mean Square Error (RMSE). In addition the MSE and RMSE values for baseline (mean observation) and for persistence are available, along with the Correlation. It is noteworthy that when MAE and ME are equal, the historical simulation is underestimating all the hydrograph, but if they are equal and of opposite signs, it overestimate the hydrograph. In general one wants the ME to be less than MAE. The deterministic verification of the historical simulation aims at providing the forecaster with the ability to discern the impacts of errors in the simulation which provided the initial condition on the ensemble performance. Higher confidence can be obtained for larger samples, which is available when 6 hourly hydrographs and stages are verified. One must be careful not to treat the full ensemble in deterministic manner. Only the historical simulation can be treated this way.

### 3. HINDCAST EXPERIMENT EMPIRICAL CUMULATIVE DENSITY PLOTS

One useful graphical representation of grouped forecasts is the cumulative density function (CDF) plot. Both empirical and probability-distribution fitted plots provide means to view the entire verification data set. Figure 11 shows the CDF plots for each of the ensembles (blue) and for the observations (bold red), for daily flows. The observation CDF represents the observation climatology and should not be compared directly to any single ensemble CDF. Similar to the plots available through the NWSRFS (espadp) tool, each CDF represent the empirical distribution of  $i$ th day forecast from all traces within all ensembles. The difference here is that multiple hindcasts are also available. This plot forms the basis for the conversion of the probabilistic forecast into binary-probability forecast, which allows the calculation of several verification measures. It must be mentioned that while the empirical distributions of forecasts and observations are used in creating the verification data set, the actual individual observations are considered as the determinant of whether the event (below/above threshold or within a range) has actually occurred. The CDF of the ensemble associated with that observations (year, day, or multiple days in case of augmented forecasts) is then used to determine the forecast probability. Therefore, one must consider the plot in figure mainly as a summary of the hindcasting experiment. Figure 12

shows the CDF but when the entire set of ensembles (i.e., 45 ensembles of 46 members each) was considered for each forecast day within the analysis period. This figure, on the other hand expresses more information regarding the similarity between the observed (climatological) and the forecast probabilities.

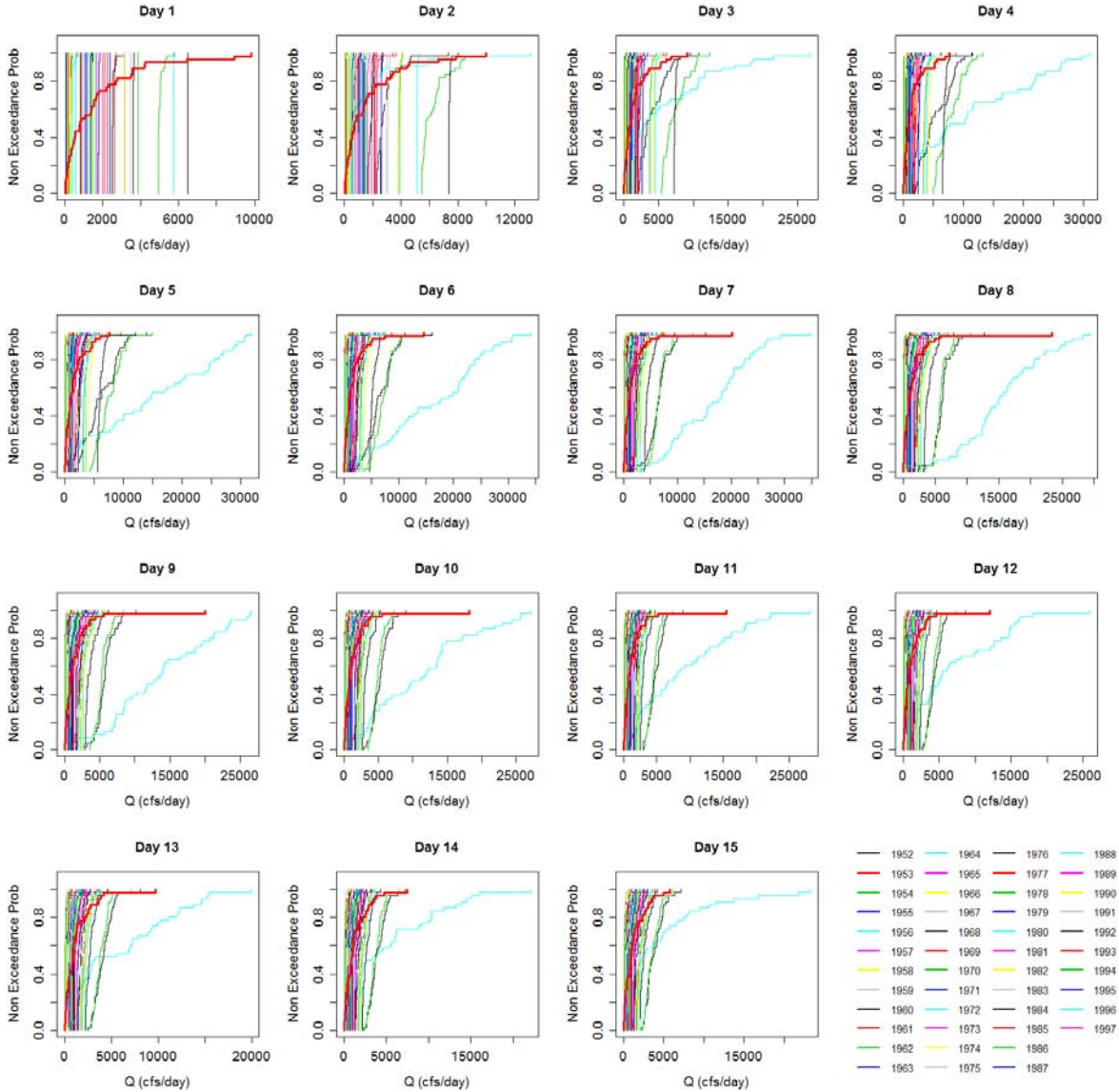


Figure 11. Empirical CDF plots for the suit of ensembles for April 1-15 representing 45 ensembles (1953-1997) with 46 traces (1952-1997) in each ensemble were used. Notice the sharpness of the ensembles' CDF plots.

*SCREENSHOT 4*

While figure 11 may not be a necessary screenshot, figure 12 below can provide good verification information to forecasters. By joining all the ensembles in the re-forecasting experiment together, and computing and visualizing their CDFs along with that of the

observations, the forecaster can envision, for each day of the forecast, the extent of the system’s ability to capture the general shape of the observation’s climatology (calibration). She/he can further discern whether the system is able to forecast events that did not occur, which can be indicated by the larger range of the forecasts CDF (discrimination). Needless to say, these are only very qualitative measures that must be accompanied by quantitative verification measures as will be seen in following sections of this report.

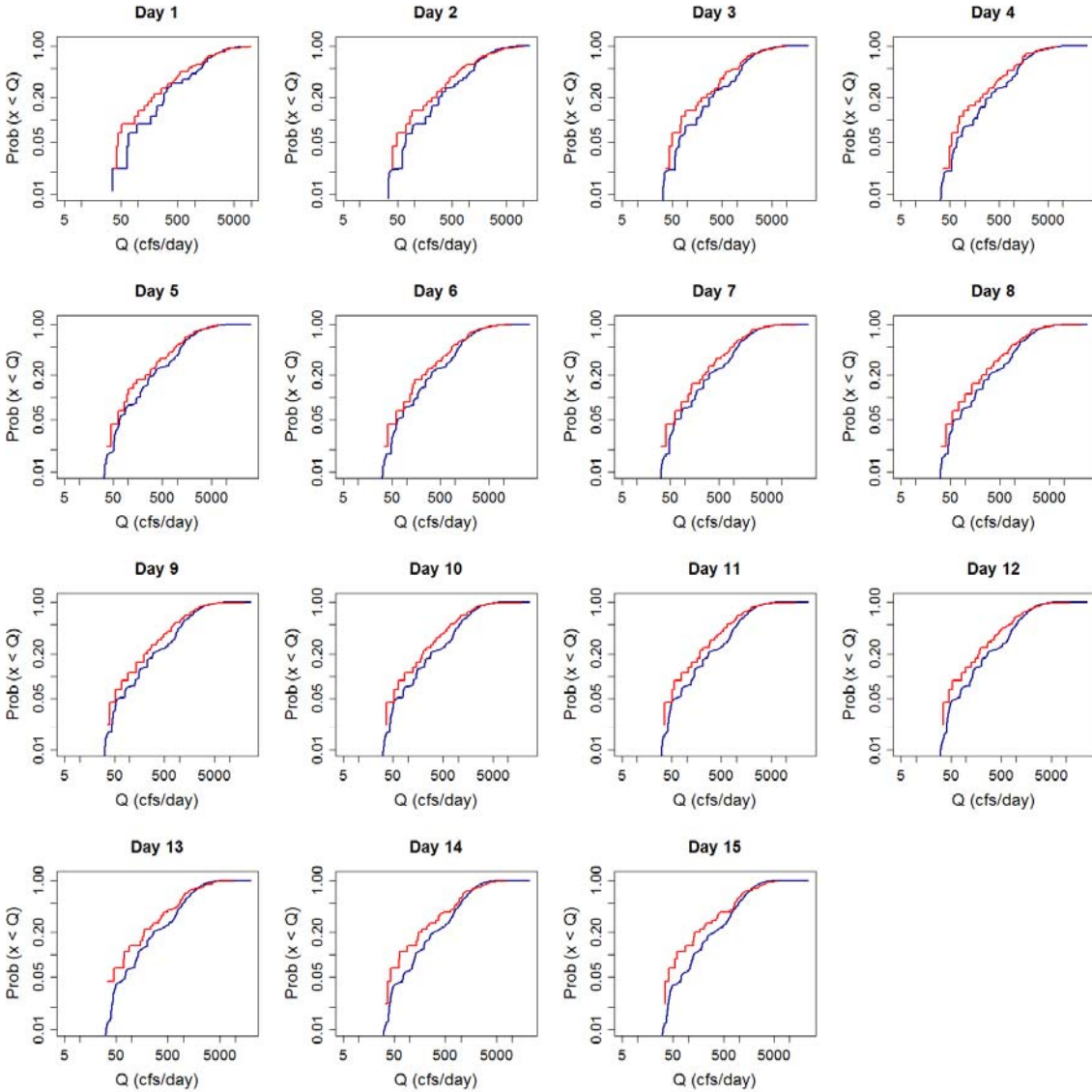
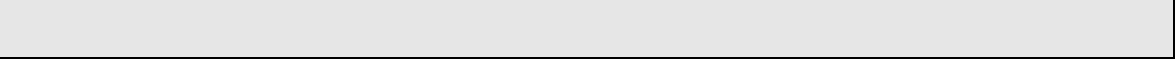


Figure 12. Empirical CDF plots for the suit of ensembles for April 1-15 representing 45 ensembles (1953-1997) with 46 traces (1952-1997) in each ensemble were used. All were considered to provide a single sample of all possible realization of forecast traces (red: observed, blue: forecast). The extended distribution of the forecast CDF relative to the observation CDF indicates the potential ability of the forecast system to forecast events that are not within the climatology.

### III.2 SAMPLE SIZE AUGMENTATION (PREDICTAND SELECTION)

While the ability to view individual forecasts is important, verification, as mentioned above, relies on statistical treatment of multiple forecasts. This means considering the entire re-forecasting data set as a single data set, with observations at each forecast point (in time) providing a statistical sample whose distribution is compared to the statistical distribution of every ensemble to generate a single forecast-observation pair. These pairs are then analyzed and verification measures are derived. Nonetheless, visual exploration of these samples can yield important information that assists the forecaster in discerning the reasons underlying given verification measures. It can also assist the forecaster in determining appropriate measures. The examples presented in section II.1 assume that probabilistic forecasts are issued for mean daily flow for each day of the 15 days (probability of flow exceeding or not exceeding a given threshold at each day). As such, the size of the verification data set for each threshold will only be 45 observation-forecast pairs. The sample size can be augmented for short term probabilistic forecasts by considering the probability of average daily flow (or discharge) exceeding/not exceeding given probability thresholds during a given interval (3,5,7,15,20, and 30 days). Such augmentation, by means of selecting different forecast variable is shown in figure 13 for the first instance of each of the 6 periods. The verification sample size increases linearly with the size of interval. The effect of such increase on the probability distributions of both observations and forecasts can be explored, and in fact, may represent the first step toward identifying the appropriate (verifiable) probabilistic forecast. Given that at daily scale, it is unlikely that the ESP forecasting system (i.e., the entire collection of ensemble CDFs), without additional information such as PQPF, will be capable of reproducing the observation climatology (reliability). Nonetheless, a qualitative indication of the forecast resolution at daily scale could be obtained by grouping all ensembles (hindcast years) into a single sample to represent the distribution of all possible realization of the daily flow. It must be mentioned that this approach assumes that the joint distribution of the forecast/observation pairs is the same for all days within the interval (Bradley et. al., 2004).

#### SCREENSHOT 5

Figure 13, below can also be used a possible screenshot from the verification system. This figure is perhaps more important for stage and high-flow forecasts. In such scenario, augmenting the sample size is a very reasonable proposition in terms of forecasting the probability of a set of threshold stages being exceeded or not-exceeded. Such figure can allow administrators to select the forecast variable based on the system's ability to provide calibration and discrimination for a given augmented period of predictions.

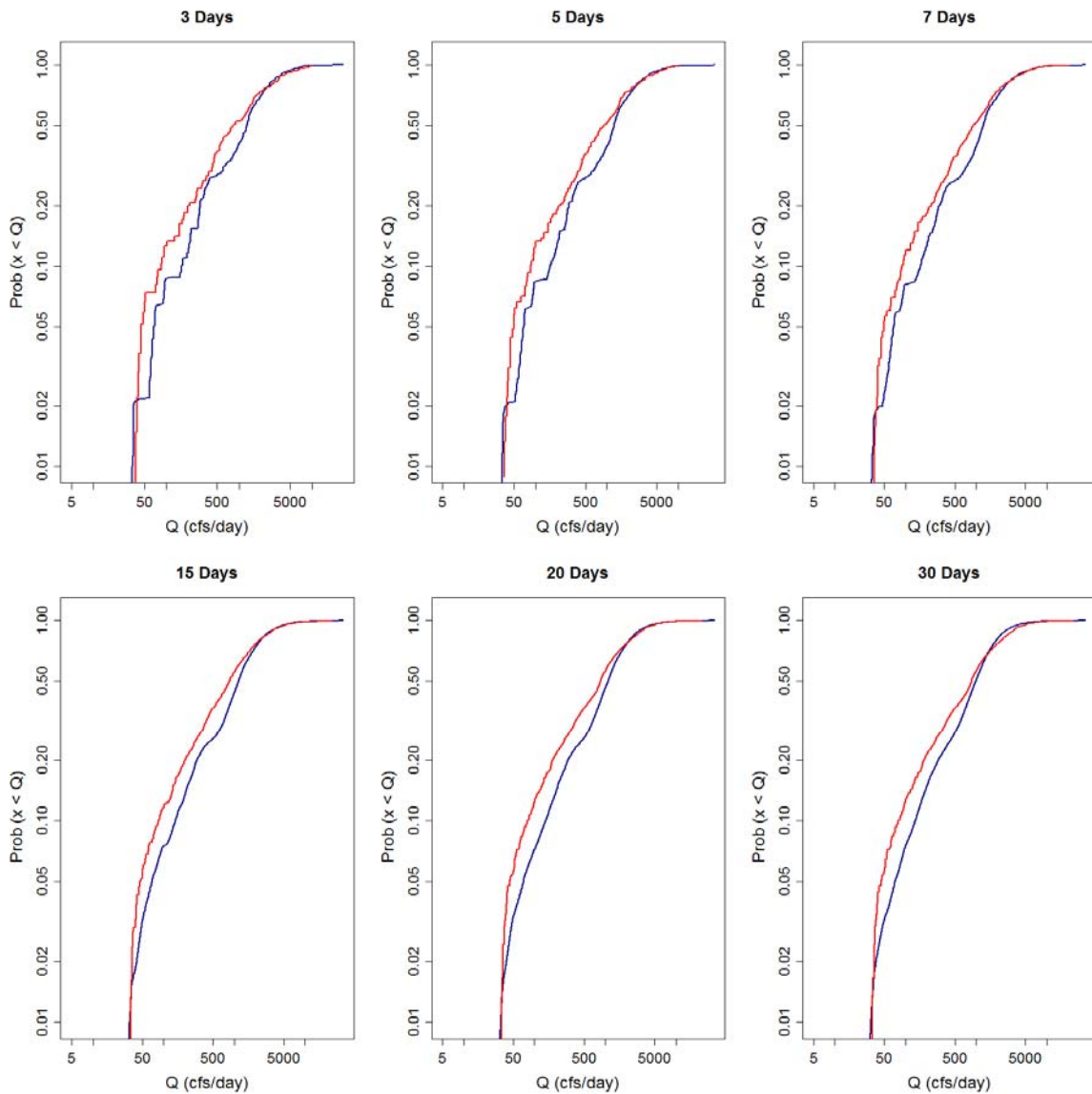


Figure 13. Empirical CDF plots for 6 temporal intervals (3,5,7,15,20, and 30) days. Notice the marked improvement in the distributions beginning with the 15 day interval. However, differences in the distribution continue to be present.

---

## IV VERIFICATION MEASURES FOR ENSEMBLE FORECASTS

---

### IV.1 INTRODUCTION

Traditionally, verification of probabilistic forecasts relied on identifying summary performance measures such as the brier score in addition to basic evaluation of the similarities and differences between the probability forecasts and the observed frequencies of forecasted events (Murphy and Winkler, 1992). These scalar measures, while helpful in

assisting decision makers in gaining information about forecast uncertainty, did not provide the forecasters and modelers with information that can be used to assess specific components of the forecasting system. A wide range of performance measures have been proposed for verification of probabilistic forecasts. However, as outlined by Toth et. al., (2003), because probabilistic forecasts can only be verified in a statistical sense, the stability and trustworthiness of any verification measure is, to a large extent, a function of the verification data set sample size (number of forecast-observation pairs). While grouping several forecasts into a single verification set increases the sample size, it reduce the amount of detailed information available on forecasts. However, for short-term hydrologic forecasts, the daily forecasts within the limited size forecasting window are less influenced by seasonal signals and grouping them will allow one to sample a wider range of meteorological forcing, all of which, can be considered equally likely in the absence of specific probabilistic precipitation and temperature forecasts. The latter is important because the limitation of the forecasting tools available to this team resulted in ESP experiments the produced ensembles which only reflect the effects of initial conditions and the climatological distribution of daily precipitation and temperatures as means to generate the probabilistic forecasts. Another important limitation of the experiments presented in this report is the fact that they are all based on hindcasts that have not been cross validated. As such, while some of the scores presented are used in diagnostic manner, they do not present, by any means, a comprehensive diagnostic verification of the NWSRFS forecasting system, a task which is not within the scope of this project.

As mentioned above, there are various measures that can be used in the context of verifying probabilistic hydrologic forecasts. Murphy and Winkler (1992) described the process of diagnostic verification of probability forecasts as analogous to that of regression residual analysis in terms of its composition of graphical and quantitative measures that attempt to explore various attributes of the probability forecasts. In this framework, the key primary attribute of probability forecasts is its statistical consistency between the forecast probabilities and the frequency of observing a given event, with two key elements: , (a) reliability, and (b): resolution. While the former measures the forecasts ability to capture basic aspects of the event's climatology, the latter measures the forecast's ability to identify events when their frequency of occurrence is different from the climatology. One must always remember that in any forecast, the forecast probability is the apriori element, while the observations are the posteriori element of the problem. As such, climatology is always viewed in the a-priori sense. In this report, we attempt to adapt their frame work, which they outlined in Murphy and Winkler, (1987), to short term probability hydrologic forecasts.

#### **IV.2 DISTRIBUTION ORIENTED (DO) MEASURES:**

Hashino et.al., (2002), and Bradley et. al. (2003 and 2004), provide an expansive treatment of ensemble forecast verification measures based on the Murphy (1997) description of DO forecast verification measures. In general, these measures require the conversion of the continuous variable probability forecasts into discrete events. With respect to ESP, these can be accomplished by identifying given thresholds of the forecast variable such as daily discharge, maximum discharge during forecast period, stage, or any desirable variable  $Q_i$ , where  $i$  denotes the current conditions (initial conditions). If one denotes this forecast variable as  $q^*$ , and identifies the probability of none-exceedence

$$f_i(q^*) = P\{Q_i \leq q^* \mid \alpha_i\}$$

where  $f_i$  is the probability forecast, and  $\alpha_i$  is the initial condition. Subsequently, define the observation variable  $o_i(q^*)$  as

$$o_i(q^*) = \begin{cases} 1 & \text{if } Q_i \leq q^* \\ 0 & \text{if } Q_i > q^* \end{cases} \quad (1)$$

A forecast verification data set for  $q^*$  can then be developed from hindcasts and climatology for the desired forecast period and for both a discrete set of thresholds.

For each threshold verification data set, the joint distribution of forecast-observation  $h(f,o)$  can be identified, which includes all the information required to assess the quality of ensemble forecasts. Murphy (1996) showed that  $h(f,o)$  can be factorized using Calibration/Refinement (CR), and Likelihood Base Ratio (LBR) as following

$$\begin{array}{ll} \text{CR Factorization} & h(f,o) = y(o \mid f)p(f) \\ \text{LBR Factorization} & h(f,o) = r(f \mid o)t(o) \end{array}$$

Where  $y$  and  $r$  are conditional distributions and  $p$  and  $t$  are marginal distributions of the forecast and observations. These factorizations allow the computation of several forecast quality measures. According to Murphy and Winkler (1987), the conditional distribution  $y(o/f)$  indicates how often different observations have occurred when the a given forecast  $f$  was issued. Given probabilistic Streamflow forecast and definition (1) for a given threshold event, the value of  $y(o = 0/f=0.85)$  indicates the frequency of threshold being actually exceeded in the observation, when the forecasted probability of non-exceedence was 0.85. Conversely,  $y(o = 1/f=0.65)$  indicates the frequency of the threshold not being exceeded when the probability of non-exceedence was about 0.65. Clearly, the ideal forecast would be one that accomplishes the following conditions:

$$\begin{array}{l} \min y(o = 1 \mid f = 0) \\ y(o = 1 \mid f) = f \quad \text{for all possible values of } f \end{array}$$

A perfectly calibrated forecast is one that satisfies  $E(o \mid f) = f$ . The marginal distribution of the forecasts  $p(f)$  illustrates the frequency distribution of forecasts for the selected threshold  $q^*$ . Sharp forecasts are those which discern certain probabilities and refined forecasts are those which cover more or less the entire range of observed frequencies. Murphy argued that the worst case scenario is that when  $p(x \mid f) = p(x)$ , which implies that the probability of event occurrence is independent of the forecast and a climatological probability is being forecasted constantly. Such forecasts lack resolution in the sense that they can not forecast events outside the climatological probabilities.

### THE RELIABILITY DIAGRAM

The CR factorization is the basis for two key diagrams (Reliability Diagram, Attributes Diagram) that are used in forecast verification. Both diagrams represent the same data. The reliability diagram (Figure 14) represents  $E(o \mid f)$  plotted against  $f$  (binned) along with a



histogram of  $p(f)$ . Following Hsu and Murphy (1986), the origin of the reliability diagram is based on the decomposition of the commonly used Brier performance measure into reliability (REL), and resolution (RES) (see Hsu and Murphy for computational details)

$$REL = E_f (\mu_{o|f} - f)^2 \quad RES = E_f (\mu_{o|f} - \mu_o)^2$$

$$BS_c = \mu_o(1 - \mu_o) : \text{Brier score for climatology}$$

$$SS = 1 - \frac{B}{B_{ref}} \quad \text{Skill score also } SS = \frac{RES - REL}{BS_c}$$

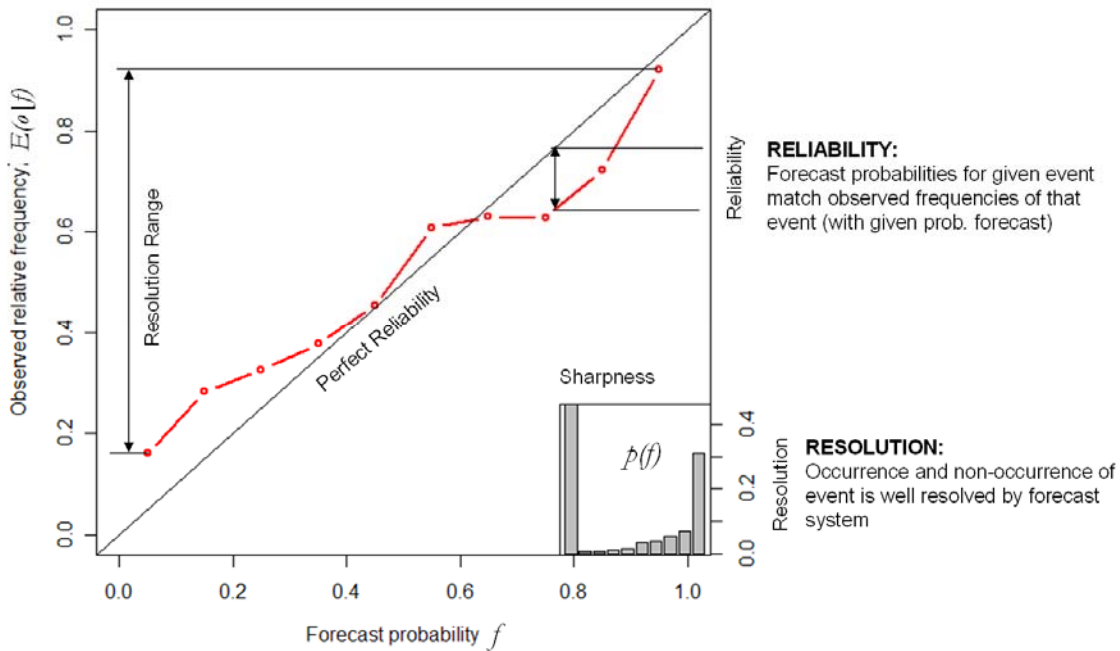


Figure 14. Reliability-Resolution Diagram. The forecast shown has reasonable reliability along some resolution. It can also be said that this forecast shows sharpness as indicated by the shape of the distribution; in this illustration, the larger majority of forecasts are in the lowermost and upper most probability ranges (sub-sample).

Interpretation of the reliability diagram is presented in the figure 15. For a given climatologically probability (dashed blue lines), the system has perfect reliability when the frequency of event occurring equals the probability forecasts equal for all probabilities. The system is under-confident, when for each forecast probability, the event occurred more frequently than the forecasted probability. Over confidence occurs when the event occurred at lower frequencies than the forecast probability. The no reliability line, which represent the climatological frequency (mean or median) frequency of the event occurring. When the diagram falls on this line, it indicates that no matter what the forecast probability was, the event has occurred with frequency equals to its climatological mean. The diagram referenced as “Anti Skill” indicate that forecasting system is assigning high probabilities to events that

occur at low probabilities and low probabilities to more likely events. Finally, interruptions within the diagram indicate that the verification sample size was too small to account for all possible probabilities.

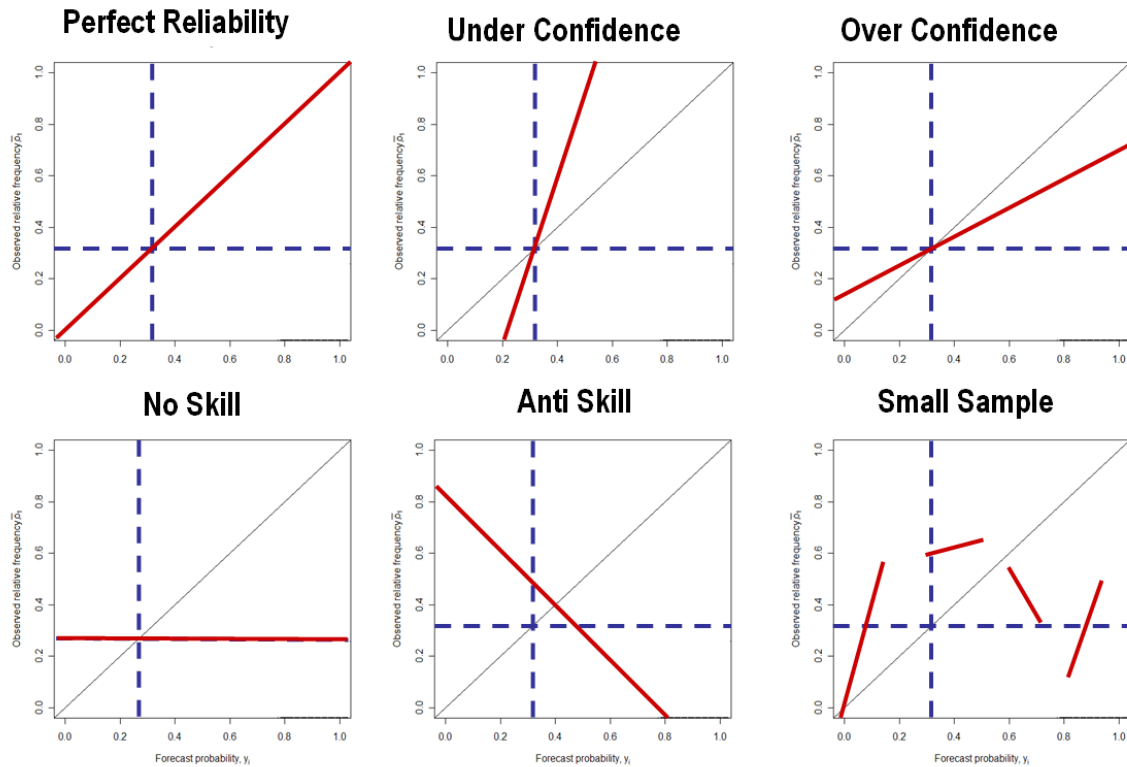


Figure 15. Interpretation of the Reliability Diagram. Considering the diagram summarizes the conditional distribution of observations given the probabilistic forecast, each point represents the expected frequency of the event occurring given the forecast probability.

Further information about the reliability diagrams could be obtained by plotting, along with the actual probabilities, the values of the Brier Score decomposition. Figure 15 illustrates these values for the selected forecast. Notice that both the local and global values of performance measures are included in said plot.

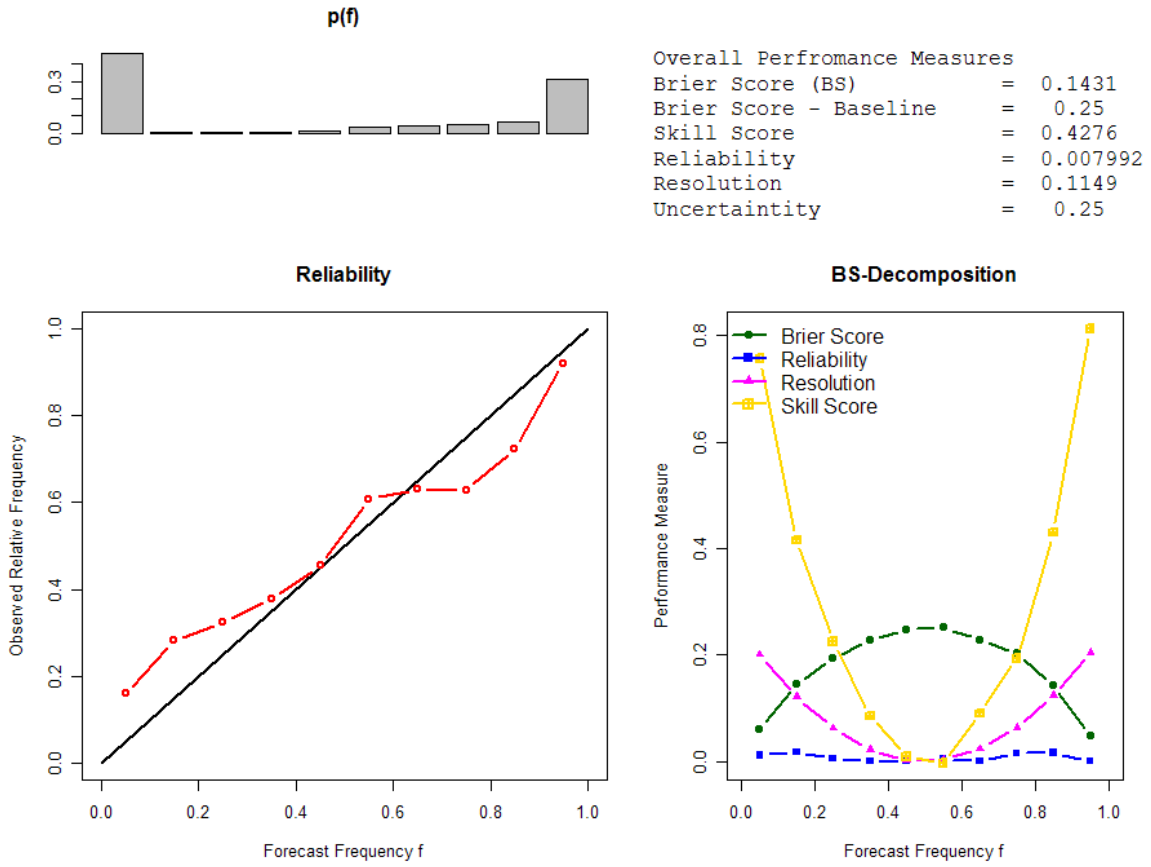


Figure 16. Reliability-Resolution diagram with decomposition of local Brier score values. The information contained in the BS-decomposition complements the information in the reliability curve. The text on the upper right shows the overall performance measures associated with this forecast. Notice skill score “gold solid squares), which is  $> 0$  for all  $f$  ranges, with skill declining at the climatological probability of the event, but increasing with increasing sharpness. This behavior is also demonstrated by the brier score (lower is better).

*SCREENSHOT 6*

Figure 16, can be utilized as possible screen to be generated by a verification system. It allows the forecaster to monitor, quantitatively, the association between various verification measures, the reliability diagram and the system’s performance. This association can be further reinforced if supported by figure 15 as a possible “help” illustration. One can envision a hierarchy of screen transitioning from simpler, data visualization screens into more complex verification measures systematically as well as within few steps. For example, forecasters familiar with verification measures can start at their selected choice but then move back to individual forecast exploration (Screens 1,2,3, and 4). And in the meantime accessing various illustrated help screens. In addition, it is anticipated that most of the screens will be interactive, which means that forecasters should select an element of each screen and proceed to drill down to more detailed exploration. It is possible that the

forecaster may only see, in the beginning the reliability-resolution diagram and the summary “overall performance” measures. Through menus the detailed BS decomposition could then be added, and the forecaster can utilize the mouse in interactive manner to move over various points on the reliability diagram and see the associated decomposition (a vertical line going through the right panel and a balloon window showing the numerical values at that forecast probability range).



A better understanding of the reliability diagram can be achieved when one considers a “bad” forecast. Figure 14 shows such a forecast, which was generated randomly (random forecast). Notice the relationship between the reliability and resolution for all probability bins. (i.e., reliability = resolution  $\rightarrow$  skill=0). Notice also the fact that the Brier score fluctuates independently of both reliability and resolution. Finally the lack of sharpness is obvious from the  $p(f)$ , which provides almost equal probability for any possible forecast. (The forecast does not say much. It is worse than simply using the climatological mean/or median of the observations ( $p(o = 1)$ )).

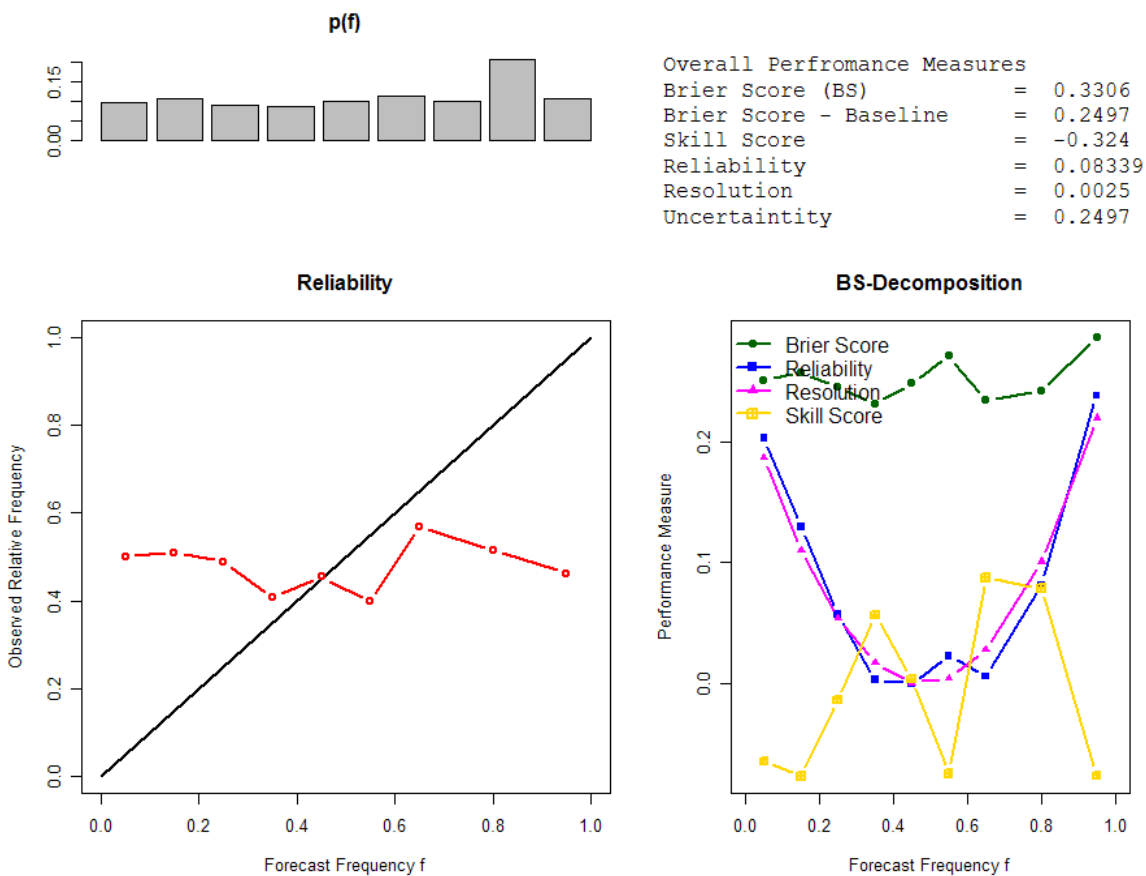


Figure 17. Reliability-Resolution diagram with decomposition of local Brier score values for a random forecast.

## THE ATTRIBUTES DIAGRAM

The second diagram that is rooted in the CR decomposition is the Attribute Diagram. The diagram adds significant amount of information to the reliability diagram, which only contains the perfect reliability line (45° line) through additional reference lines. While its original formulation (Hsu and Murphy) included many reference lines, in practice, the four most commonly used lines are (a) perfect reliability, (b) no-skill, (c) no-resolution and (d) the vertical line where the forecast probability equals the climatological mean of the observation or median for variables displaying high value of the skewness coefficient. As seen in figure 18, these lines define a shaded region, outside of which the forecasts will have no skill. An additional feature of the attribute diagram is the inclusion of textual labels representing the frequency of forecasts within each of the probability ranges considered, which provides the same information as those in the reliability-resolution diagram. Figure 19 shows the attribute diagram for the random forecast discussed above, along with the various performance measures superimposed (y axis on left). Clearly, the value of the attribute diagram is made evident (see figure caption).

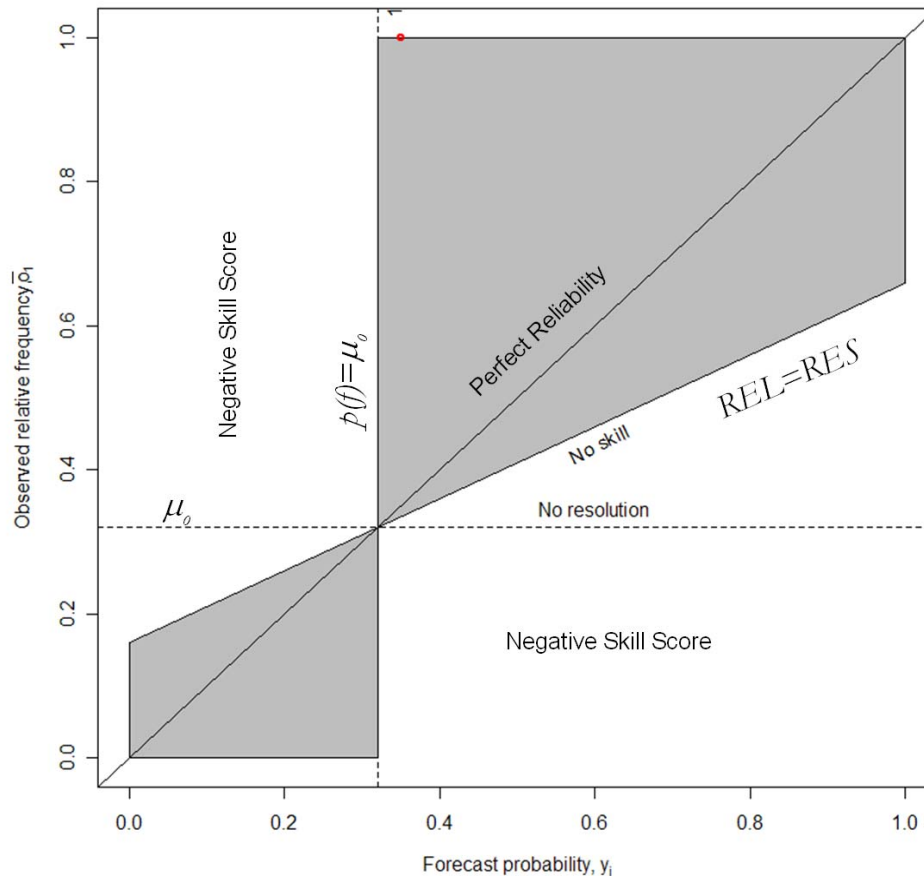


Figure 18. Major regions of the attribute diagram. Notice the addition of the regions defining negative skill score, which are associated with “Anti-Skill” (Figure 15). Also notice the gray region defining reliable forecasts and separating the regions of acceptable “Over and Under Confidence”. Although it seems more complicated, the attribute diagram adds the interpretation of the reliability diagram and provides the resolution histogram in numerical values (See Figure 18 below)

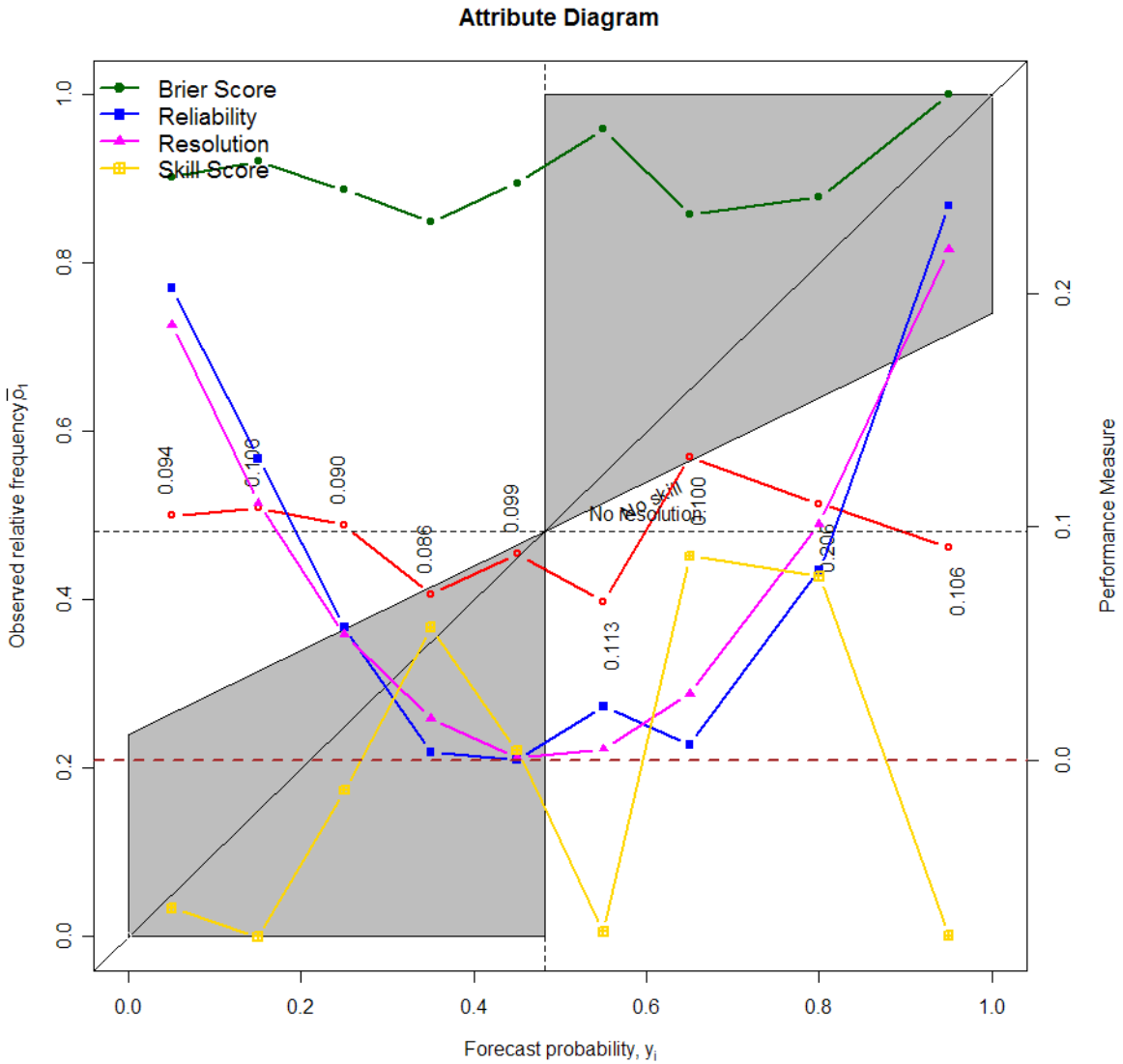


Figure 19. Attribute diagram for random forecast. Performance measures, which are not part of the diagram are superimposed with right value axis to illustrate the regions of the reliability diagram. Notice the presence of resolution information (numbers representing the histogram of  $(\bar{y}_i)$ ). Also notice that for points of the reliability plot (red line and open circles), which fall outside of the shaded region, the corresponding values of the skill score are negative. Notice also that points where the reliability (blue) and resolution (magenta) intersect fall on the no skill line with the actual frequency curve (red) being very near the no resolution line.

### IV.3 SPECIFIC ESP-CONSIDERATION

With respect to ensemble generated probabilistic forecasts, minor deviations from the diagonal “perfect reliability” line of for a given verifying data (observation) are not essentially due to lack of forecast performance in the probabilistic sense. Murphy and Winkler (1992)

and Toth et. al (2003) argue that by randomly replacing the verifying analysis (observation) with one ensemble trace, one could identify the attribute of a “near perfect” forecast given the forecasting system. From ESP point of view, instead of selecting a random trace, the selection of the “Historical Simulation” as a verifying analysis provides the basis to determine the best possible performance of the system, while eliminating the effects of model errors on the verification. By comparing the two verifications (observation based vs historical simulation), the forecasters as well as the administrators can identify whether model improvements are likely to substantially improve the forecasting system. Figure 20 shows such a comparison. In the figure, the attribute diagram shows both historical simulation and observation attributes (center bottom panel). The resolution histogram (center top panel) and the decomposition of Brier scores for both historical simulation (left bottom panel) and observation (right bottom panel) as well as the associated overall performance measures (left and right of top panel)

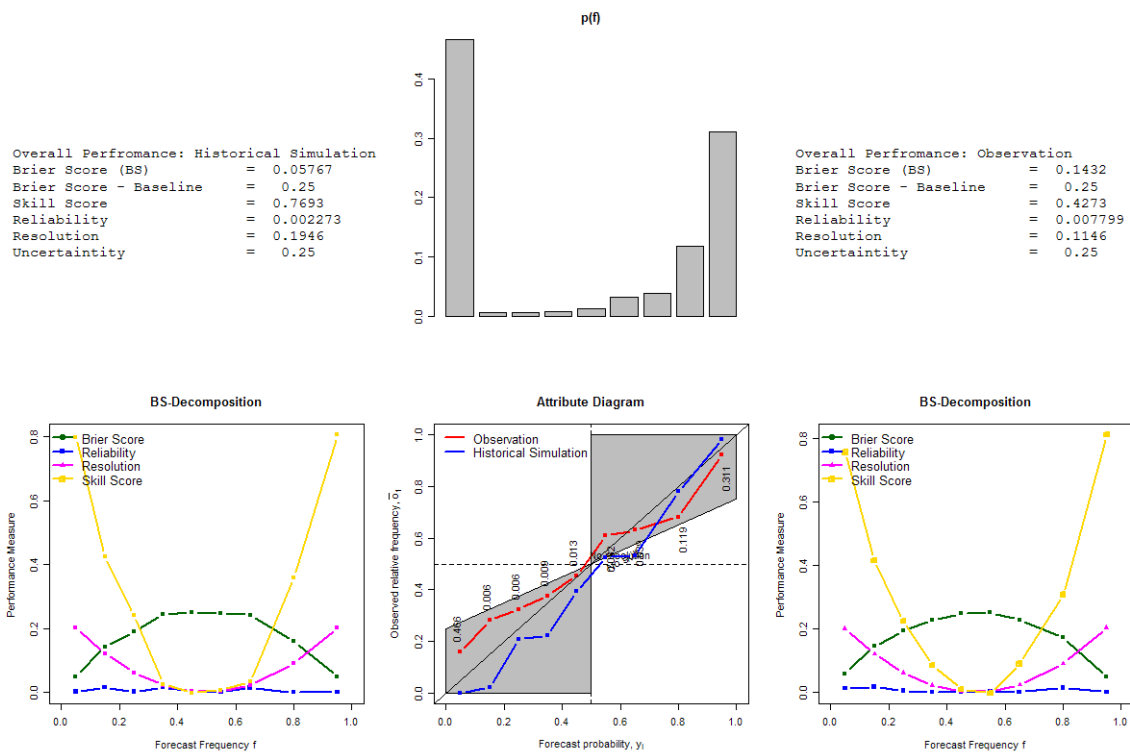


Figure 20. Referencing attribute diagram with historical simulations for April 1-15 (blue line in the attribute diagram) in order to discern forecasting system performance. In this case, both forecasts and historical simulations were generated using the NWSRFS, but the verification data set was augmented by considering multiple probability thresholds (0.1,0.2,0.3,.....0.9), which explains the near 0.5 mean of observation. Examples associated with individual probability thresholds will be presented later. Notice the better values of Brier Score, and Skill Score when historical (top-left table) simulation is considered. However, notice the reasonable performance compared to observations as well.



## SCREENSHOT 7

Figure 20 also represents an alternative verification system screen to screenshot 5. The main benefit of this screen would be to provide the means for better interpretation of the reliability diagram by using the attribute diagram instead, which provides the gray range delineation of the reliability and skills regions. Again, the screen can be made interactive with increasing level of complexity as the forecaster/administrator goes through the process of drill-down through more details. It must also be mentioned that this screenshot must be available at every threshold probability range or in case of stage and peakflow forecasts, at every critical stage and discharge chosen by the forecaster. Comments received during our verification workshop seminar indicate that such screen may be a little too complex for verification/validation objectives. While we agree with these comments on principle, we believe that such screens must be available through any forecast verification to provide for detailed verification of both probabilistic and deterministic forecasts. The historical simulation may be replaced with a simulation based on optimized model stated.(please see recommendation section)

### IV.3 EXAMPLE

Figure 21 below, shows the detailed attributes plots for 9 different thresholds. All plots were generated for the case study described above (April 1-15, 45 ensembles with 46 traces in each ensemble). It must be mentioned that these figures could also be replaced, for a given threshold, by an evolution of forecast attributes with verification of recent forecasts. Furthermore, figures 22 and 23, show the Brier score decompositions for all thresholds and for observation and historical simulations. Therefore, all can provide detailed information to forecasters at a very substantive drill-down level. Yet, at the same time, they provide a full summary of the entire hindcasting experiment.

## SCREENSHOTS 8, 9, and 10

Figure 21, when provided through the verification system allows the forecaster/administrator to visualize the attributes of forecasts for all thresholds selected during the verification. One can also switch between figure 21 and figures 22, and 23 to obtain more details on the entire verification data set both for observation and historical simulation. The ability to visualize the scores, when the ensemble is compared against the historical simulation provides the “best” possible performance of the system. A help screen that includes figure 15 can guide the forecaster in interpreting the diagrams appropriately.

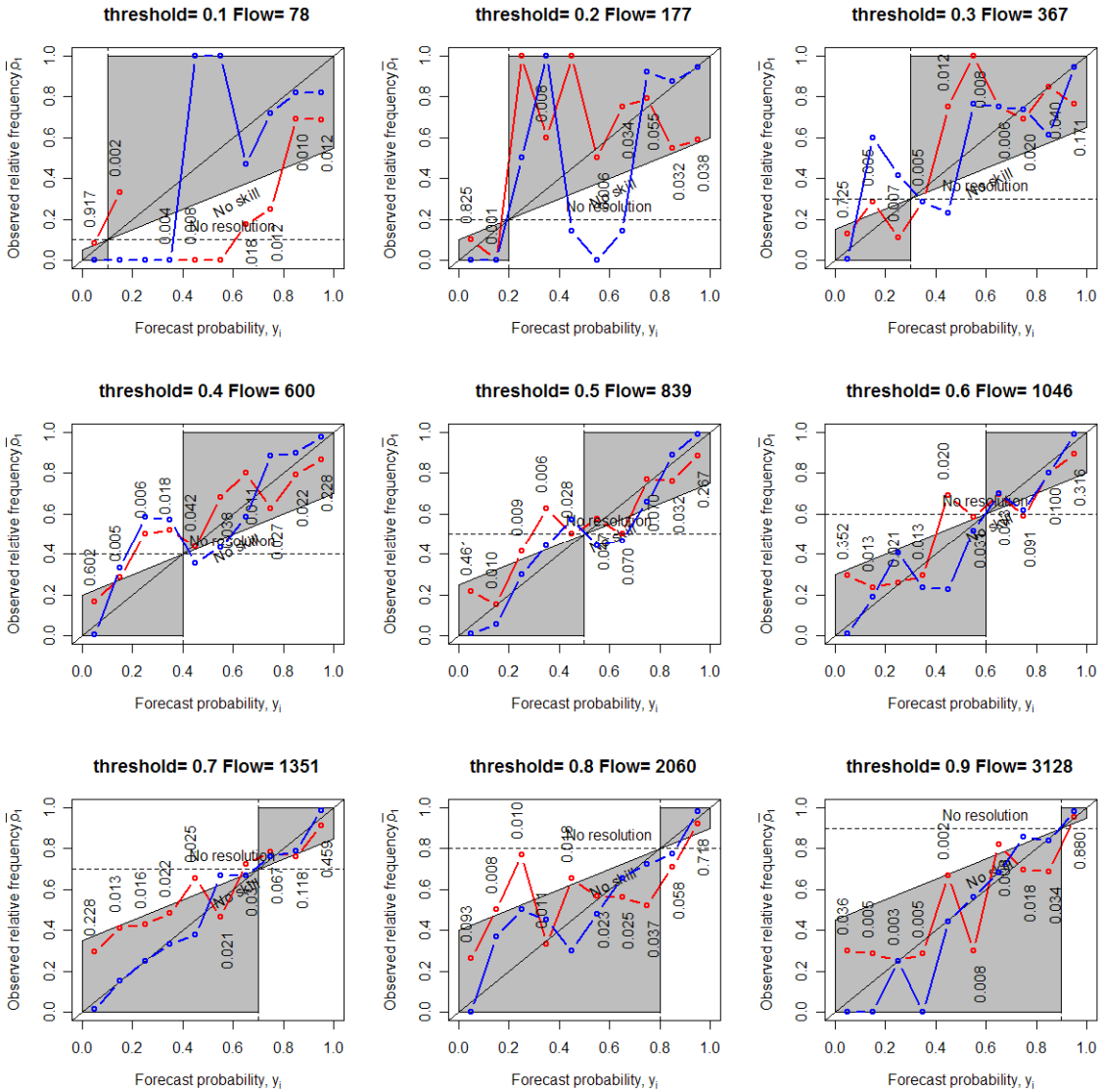


Figure 21. Sample exploration of forecast attribute for various thresholds. Notice the evolution of the forecast attribute with lower probability thresholds lacking skills and improved skills for higher probability thresholds. Two thresholds are notable, particularly when observation-based verification is compared with historical simulation verification. These are 0.5 and 0.7. However, at higher probability threshold (starting with 0.4) the reliability diagram shows a reasonable reliability. One must recall that this diagram is based on augmented sample of 15 days.

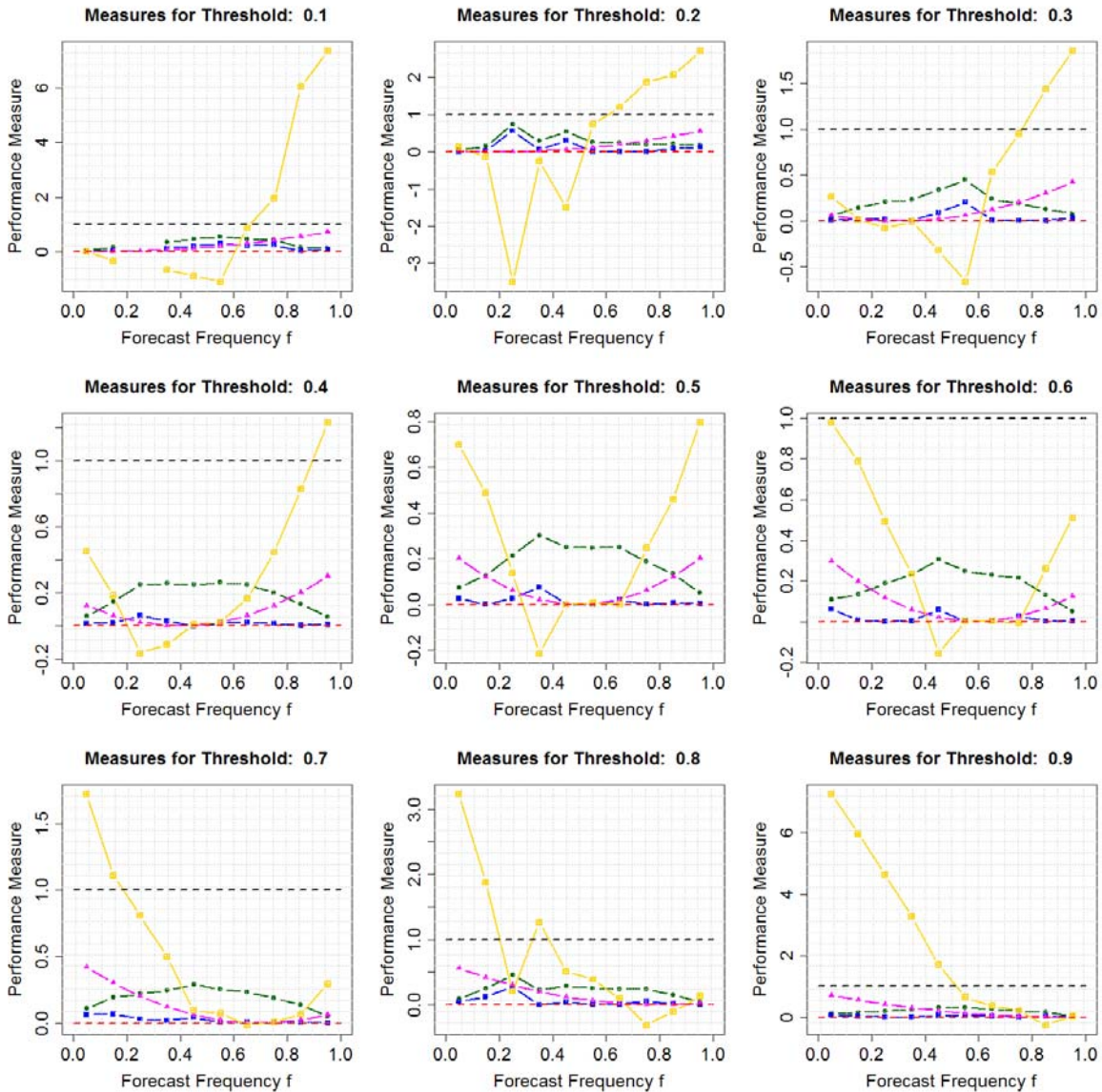


Figure 22. Sample exploration of forecast skill and verification measures for all considered thresholds when forecasts are verified against observation. The lines drawn are similar to those in figure 20, which include the Brier Score (green), reliability (blue), resolution (magenta), and skill score (gold). The dashed red and black lines represent the (0,1) interval which indicate reasonable range of some of these measures. Notice the interruptions caused by the smaller data set, or the inability of the forecasting system to address all possible probability ranges particularly at threshold=0.1. Also, notice the improved performance at threshold probabilities 0.8 and 0.9, indicating system's ability to forecast events with low non-exceedence probabilities. This figure is a supplementary screenshot to figure 21.

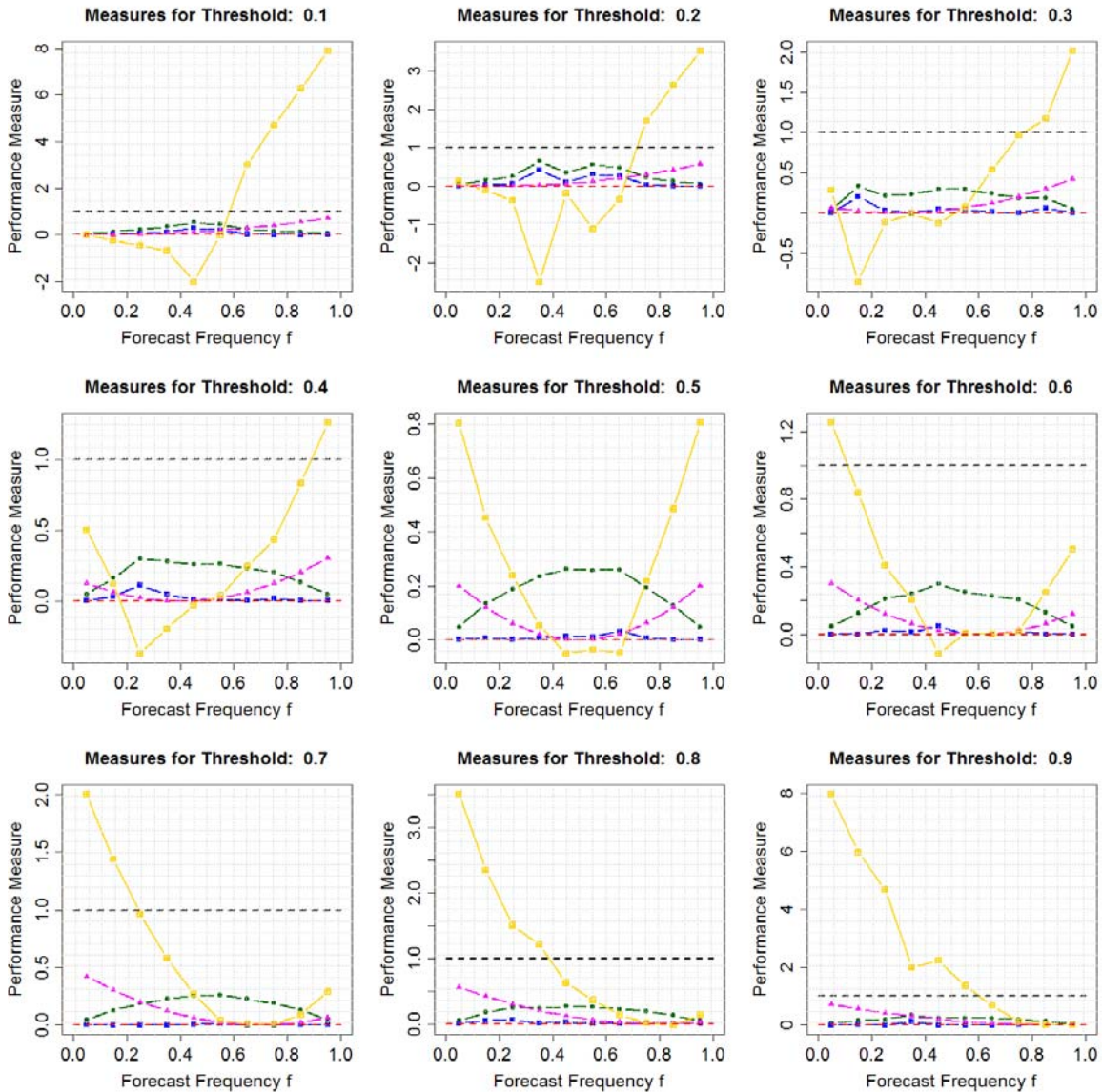


Figure 23. Sample exploration of forecast skill and verification measures for all considered thresholds when forecasts are verified against historical simulations. The lines drawn are similar to those in figure 20, which include the Brier Score (green), reliability (blue), resolution (magenta), and skill score (gold). The dashed red and black lines represent the (0,1) interval which indicate reasonable range of some of these measures. Notice the lower number of interruptions in comparison to figure 22. Also, notice that in most cases, the system verifies better against historical simulations than against observations. This calls for incorporating alternative methods (e.g., optimized model states) in the generation of hindcast as well as ESP forecasts. This figure is also a supplementary screenshot to figure 21.

## THE ROC CURVE

Figure 24 is similar to figure 20 for a given threshold, but the resolution diagram has been replaced by the summary overall performance measures, and two new diagrams are shown on the top left and right panels representing ROC (Receiver Operating Characteristic) curves for both historical simulation and observations. ROC curves were originally developed to assess the performance of signal detection classifiers. The key objective of ROC curve is to visualize the trade off between hit rate and false alarm rates (Egan, 1975). ROCs are based on contingency tables and on the LBR factorization, which also form the basis for the classical Brier score. For given a forecast, the observations of an event, which could be defined by probability thresholds (non-exceedence or exceedence), take the value of 0,1 based on whether it occurred or not. The ensemble based probabilities are then treated as scores assigned to forecasting the event. An observation/forecast contingency table (dichotomous) is determined by specifying the decision threshold (e.g. changing the probability level at which an event is forecasted to have occurred or not). Each change of the probability threshold results in a new contingency table, which is then translated into a point on the ROC curve. Figure 25, shows the main characteristics of an ROC curve. Thresholds, also known as cutoff, increase from the (1,1) corner, which represents the scenario when a positive forecast is always issued. The (0,0) corner is associated with a forecast that is never positive. The diagonal represent the case when FAR = HR, or in other words, a random forecast.

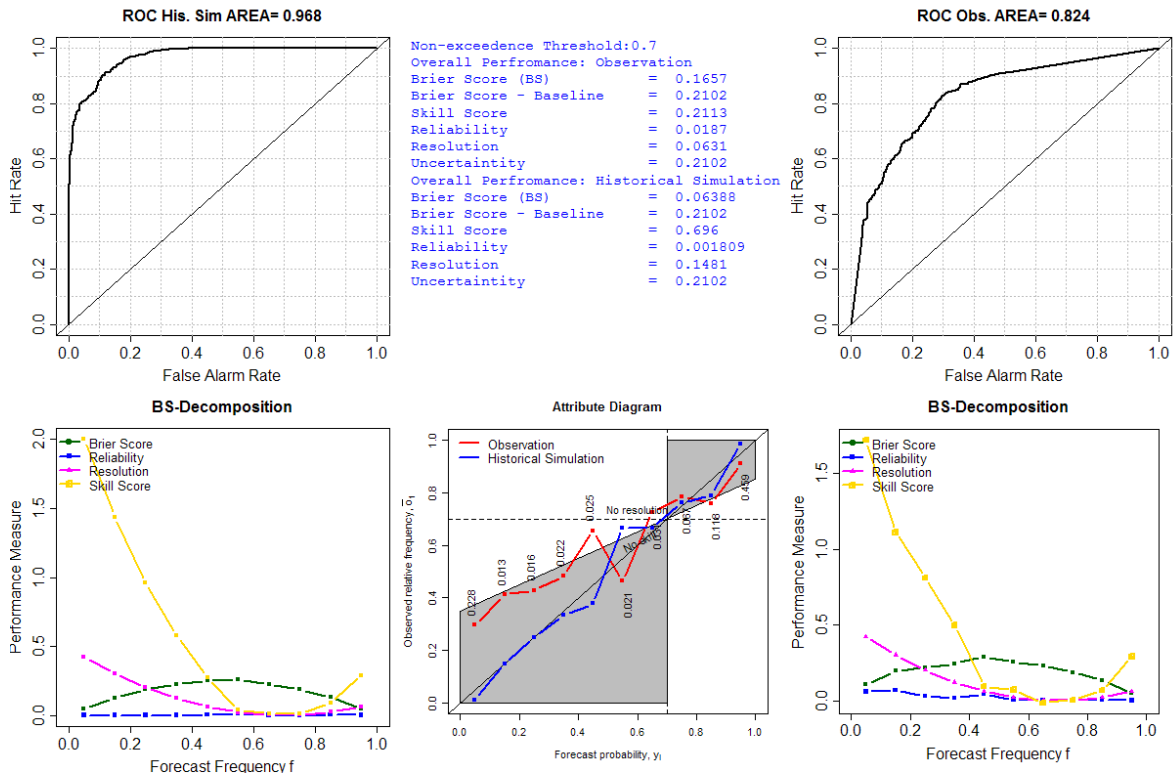


Figure 24. Sample detailed exploration of forecast attribute for a given threshold (0.7). The two ROC (Receiver Operating Characteristics) curves are based on the classical contingency table.



		Observation	
		p	n
Forecast	Y	True Positive	False Positive
	N	False Negative	True Negative
Total		P	N

False Positive rate=FP/N = False Alarm Rate  
 True Positive Rate= TP/P = Hit Rate  
 Precision = TP/(TP+FP) = Positive Predictive value  
 Accuracy= (TP+TN)/(P+N)  
 Specificity = TN/(FP+TN)

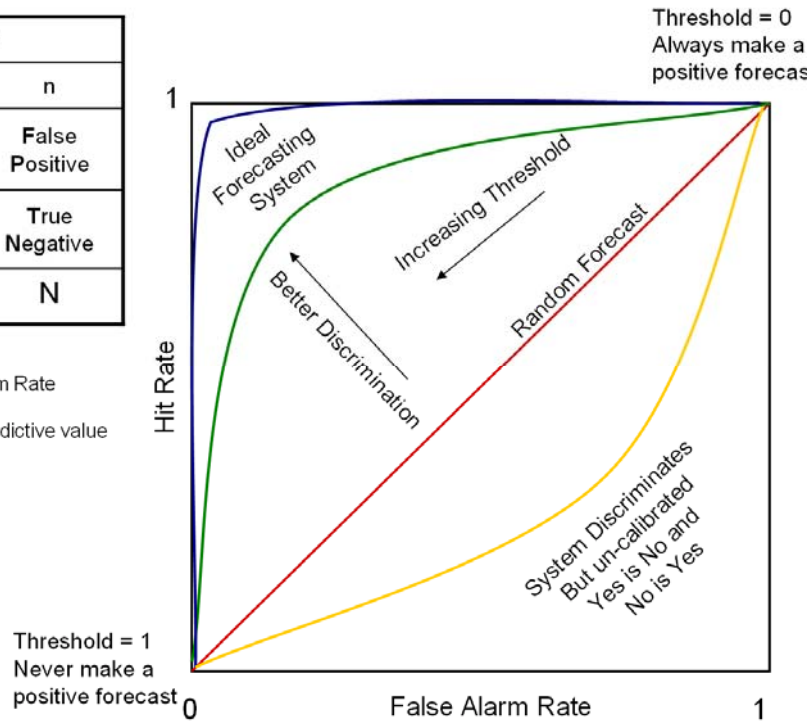


Figure 25. Construction of ROC curve. In practice, the ROC curve will not be as smooth as those shown in the figure, but better performance is associated with convex ROCs. Poorer discrimination is associated with ROC curves that alternate frequently between convex and concave shape, with both large and small concavities present in the overall shape. In addition to the visual interpretation of ROC, it is possible to also use the Area under the ROC curve as a measure of performance. When ROC Area < 0.5 the system has anti-discrimination (associated with the gold line). When the area = 0.5, the system is no better than a random forecast. As the area increases reaching 1, the system has better discrimination.

ROC curves are a very valuable verification tool, particularly when the verification sample size is small, or when the forecaster needs to make probabilistic forecasts for low frequency events (e.g., flood stage). For example, consider the probabilistic forecast of the event with the 0.7 climatological nonexceedence probability occurring during the period April 1-7. The forecast attribute diagram, shown in Figure 26 for both historical simulation and observations gives the forecaster the impression that her/his forecast is not reliable. However, when considering the ROC curves, it seems that within the possible ranges of cutoffs, the forecasting system provides a reasonable discrimination.

The above-described measures of performance provide substantial information regarding the forecasting system. As mentioned repeatedly in this report, there are numerous verification measures, but the ones described herein are specifically tailored to ensemble probabilistic

forecasts. Other measures were discussed in details by Bradley et al. (2002), In addition, the above-referenced verification team report provides a well thought list of deterministic and probabilistic measures. We attempted to focus on few but powerful measures, particularly those that can be visualized, explained, and above all interpreted within a short period of time. In the next section, we attempt, based on our communications with several forecasters to consider the key operational requirements for real-time short term probabilistic forecast verification approach. Needless to say, even in realtime, probabilistic forecast verification is always a diagnostic affair. It is not expected that the forecaster will adjust the hydrologic model in realtime based on the diagnostic verification. But it may be possible to adjust other factors including the choice of weights, probability distributions, and verification data sets.

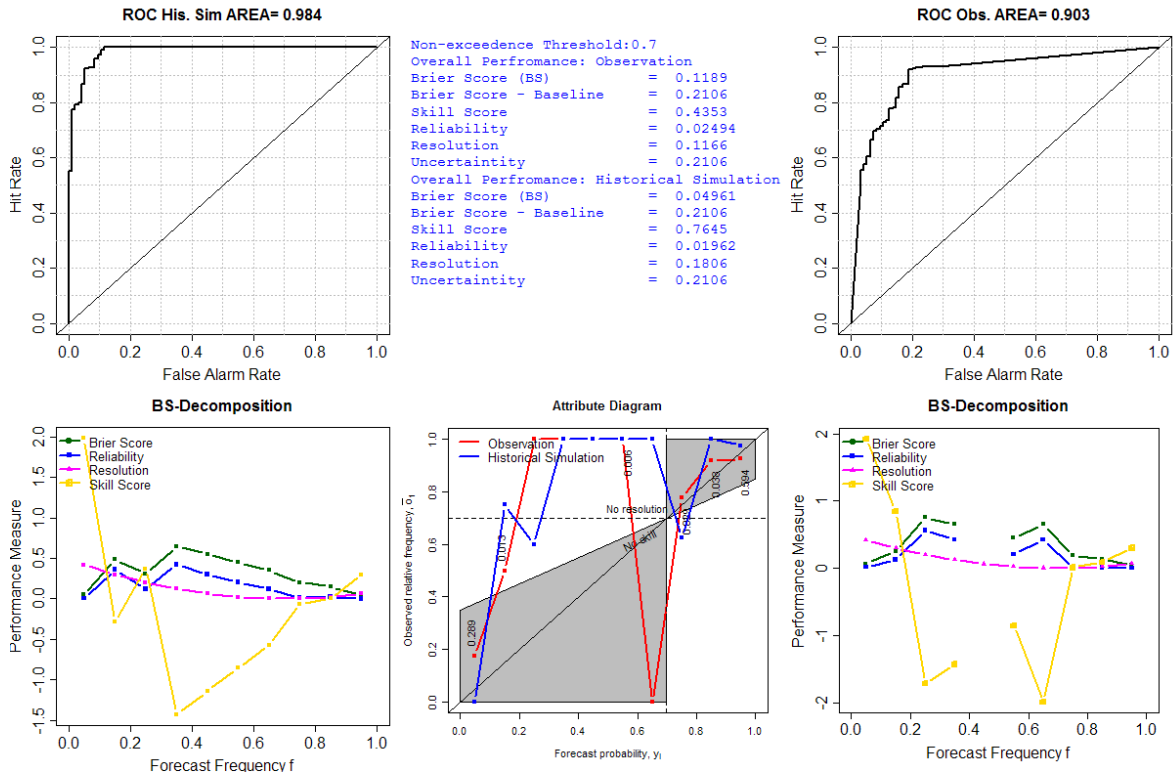


Figure 26. ROC curves can provide valuable verification information when the sample size is small, or when the attribute diagram is not very clear. The above combination represents a thorough, yet, reasonably easy to interpret verification. Forecast is the 0.7 probability of non-exceedence for April 1-7. Notice that both possible verification data sets show reasonable discrimination. The observation shows more concavities in the diagram, but the ROC is within appropriate range. Notice also the discontinuity in the reliability, Brier score, skill score, and resolution curves. These are related both to small sample size as well as failure to account for initial conditions appropriately during the hindcast procedure (See figures 7 and 8).



## SCREENSHOTS 11 and 12

Figures 24 and 26, which represent the same information, form the basis for the most detailed and comprehensive visualization of formal verification metrics. The screenshots include the attribute diagrams for the forecast being verified against both observation and historical simulation, the decomposition of the Brier score both in overall (across all forecast probabilities for the selected threshold), as well as the detailed probability forecast values of the brier score and its decomposition (Calibration/Refinement). The likelihood base ratio decomposition is also present in the form of the ROC plots for both observed and historical simulation based verification. Forecasters and administrator are likely to utilize this part of the future forecast verification system after they become familiar with the simpler screens described above.

---

## V DETERMINISTIC VERIFICATION MEASURES

---

### V.2 VERIFICATION FOR CONTINUOUS VARIABLES

Many times in this report we mentioned the need to verify using both observations and historical simulations. This is mainly because of the significance of historical simulations initializing ESP hindcast experiments. In addition, during several discussions with various RFCs and with members of the NWS-verification team, the need to illustrate some deterministic forecast verification measures were discussed. In this section we provide example screen shots for deterministic forecast verification measures. We use the historical simulation in this context because we believe that it is important to verify the historical simulation as a stand alone deterministic forecast. This was illustrated for a given hindcast year in figure 10, and screenshot 3. In figure 27, we consider all forecast years and compute verification statistics in manner similar to that of figure 10. This represents the two forecast and observations as continuous variables. As seen in the figure, Notice that for the very high flow year, the forecast (historical simulation) performed well as well as for other years with high flow. It is possible that focusing on the RMSE during model calibration may have contributed to that. However, from ESP hindcast generation point of view, capturing the initial conditions and model states in both high and low flow years is very important.

### SCREENSHOT 13

Figure 27 represents a possible screenshot for the deterministic forecast verification of the a verification system. The figure includes 4 panels. A summary of formal verification panel (top-left), a time-series plot of observed and predicted (top-right), a scatter-gram with the size of dots dependent on the relative magnitude of error, and finally a standard conditional quantile diagram. Although the figure helps the forecaster/administrator in determining the

level of errors in the historical simulation in ESP context, it is valid for any deterministic forecast.

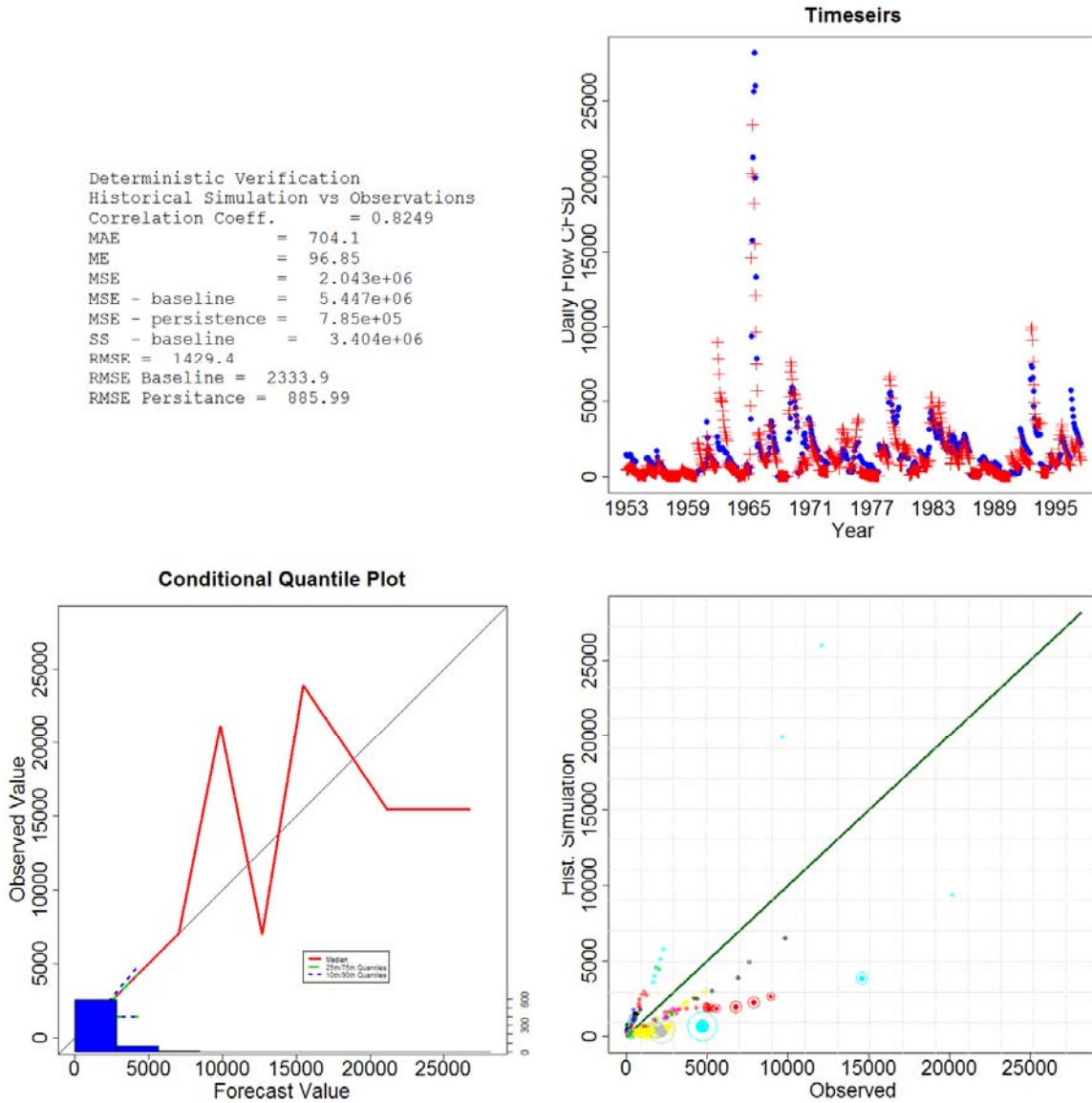


Figure 27. Possible graphical representation of deterministic verification component of the verification system. In this case, the historical simulations for April 1-15 are all verified against their paired observations. The top-left panel includes many of the verification measures proposed by the NWS-verification team. The top right panel shows the observation (red-dots) and historical simulation (blue dots) for each of the forecast years. The bottom-right panel is a revised version of the classical scatter-plot with the size of each “bubble” being determined by the relative magnitude of the error ( $Abs(Obs-Pred)/Obs$ ). Notice that for the very high flow

year, the forecast (historical simulation) performed well. As well as for other years with high flow. It is possible that focusing on the RMSE during model calibration may have contributed to that.

## V.2 VERIFICATION OF CATEGORICAL FORECASTS

In many cases, both deterministic forecasts and probabilistic forecasts are issued in the form of categorical forecasts. With respect to multi-categories probabilistic forecasts, the verification approach is identical to the binary case, with thresholds being replaced by categories. Reliability, attributes, and ROC diagrams along with Brier score decomposition remain valid. However, when both forecasts and observations are strictly categorical (i.e., event occurs or not, or forecast stage to exceed certain level or be within levels), other scores become available. These include the the Heidke Skill Score, the Pierce Skill Score, the Gerrity Score, and the Threat Skill score along with the FAR, and Hit Rate.

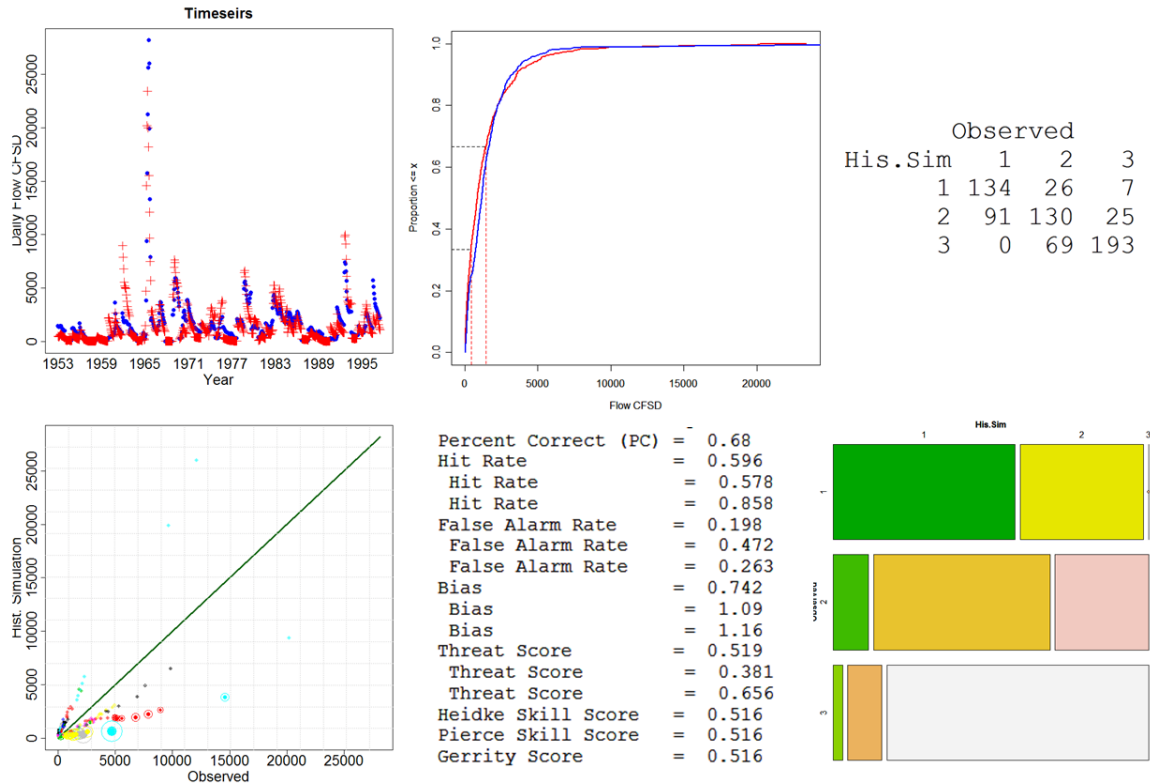


Figure 28. Categorical verification measures for deterministic forecast (April 1:15). Notice the benefit of having the structured contingency table along with the scores. Notice also the correspondence between the time-series plot and the fact that the hit rates for the higher flows is reasonably high (0.858).

Figure 28 represents a possible screenshot for the deterministic/categorical forecast verification of the a verification system. The figure includes most of the relevant metrics considered by the NWS-Verification Team. The combination of graphical and textual representations of the contingency table should allow the forecaster/administrator to better interpret the skill scores available in the panels. In addition having both the time-series and the scatter plot will provide the detailed (drill down) information many forecasters are accustomed to.

---

## VI OPERATIONAL CONSIDERATIONS

---

Welles (2005) emphasized that the objective of administrative verification of deterministic river stage forecasts is to determine:

1. How does the performance of the actual forecasts compare to the performance of persistence forecasts?
2. How does the forecast performance change with lead time
3. How does the forecast skill change with time

Answering these questions is critical to the mission of any operational forecasting agency. It requires a long-term archive of actual forecasts. Note that ESP reforecasts may address (1) and (2). Alas, no long-term archive of actual ESP forecasts exist for several reasons including:

- (1) Recent implementation and continuing evolution of ESP procedures at RFCs
- (2) Lack of archival procedures of actual ESP forecasts in their numeric (ensemble) format.

Clearly, in the near-term, any serious verification of ESP forecast will rely heavily on conducting hindcasts similar to those presented in this report. In fact, the NWS verification team has clearly identified the need for both a verification system and thorough system wide verification capability. They provided detailed description of the major requirements of such system, which includes both deterministic and probabilistic verification capabilities. The Report also identifies a suite of verification measures and our team will work with the verification team on providing graphical examples of some these measures in order to emphasize the needs for data exploration capacities within the anticipated verification system. While the data reported in this report is by no means comprehensive, or sufficient, it points to a serious issue that must be addressed in developing operational procedures for real-time verification of short-term ESP forecasts. In addition, until the utilization of PQPFs

(ensemble) becomes a standard operational procedure in ESP, short-term forecast performance as a function of lead times is of less concern than that of selecting the appropriate predictand, which is verifiable. A second complicating factor is the apparent disconnection between the hindcasting procedure and tools available and the actual operational utilization of ESP tools. On the one hand, some modifications implemented by the forecaster are by default included in her/his ESP forecasts. The hindcasting procedure does not allow for the historical record (even if it exists) of these MODS to be incorporated. Therefore errors in model carryover states that may have been corrected by MODS may exist in the initial conditions of the hindcasts, causing bias in the distribution of the resulting ensemble that would not have been present in the real time ESP.

Therefore, real-time operational verification of short term probabilistic forecasts must take into account that the verifying hindcast data is not necessarily an accurate representation of the system or of the forecaster's skill. The size of the verifying data set must then be augmented through procedures such as focusing on flood stage (action, flood, major flood) events thresholds, which allows for a suite of forecast points within a given forecasting group to be lumped in one verifying data set. Selection of verifying data set needs to take into account factors such as:

1. Similarity with current conditions (e.g., snow pack, meteorological conditions, current stage). For example, the forecaster can widen the verification data by considering all forecasts issued on days with similar, as well as hindcasts initialized on such days. This will replace the standard hindcasting procedure, which relies on calendar dates by a system that has the ability to query the historical data and select larger sample, but within reasonable seasonal and short-term conditions.
2. Skills or exploratory analysis of recent ESP forecasts. This will allow the forecaster to identify, within reasonable time window, the discrepancies between recently issued ESP forecasts and their corresponding observations.

As mentioned above, it is not expected that the forecaster will adjust the hydrologic model in realtime based on the exploratory analysis of recent forecasts or even based on diagnostic verification of hindcasts. However, a verification procedure that accounts for current conditions will allow the forecaster to better identify appropriate probability distributions, as well as probability ranges. Figure 22 shows a conceptual vision of the relationship between verification system and the current operational NWSRFS.

As mentioned above, the current implementation of hindcast generation does not account for modifications and adjustments made by forecasters. It is our belief that re-forecasts will always be an integral element of hydrologic verification system. And as such, alternative approaches to establishing model initial conditions must be identified that attempts to capture, to the most possible extent, the effects of forecasters' judgment on the initial conditions of hindcasting experiments. This can be accomplished through varieties of techniques including (a) integrating data assimilation or filtering tools into the hindcast generation tool, (b) creating an alternative "historical simulation" with optimized initial conditions, and (c) maintaining an archive of "historical" optimized model states that can be generated during the re-calibration process, among many. This is consistent with the

conclusions arrived at by the NWS verification team (Gabrielsen et. al, 2006 Draft document).

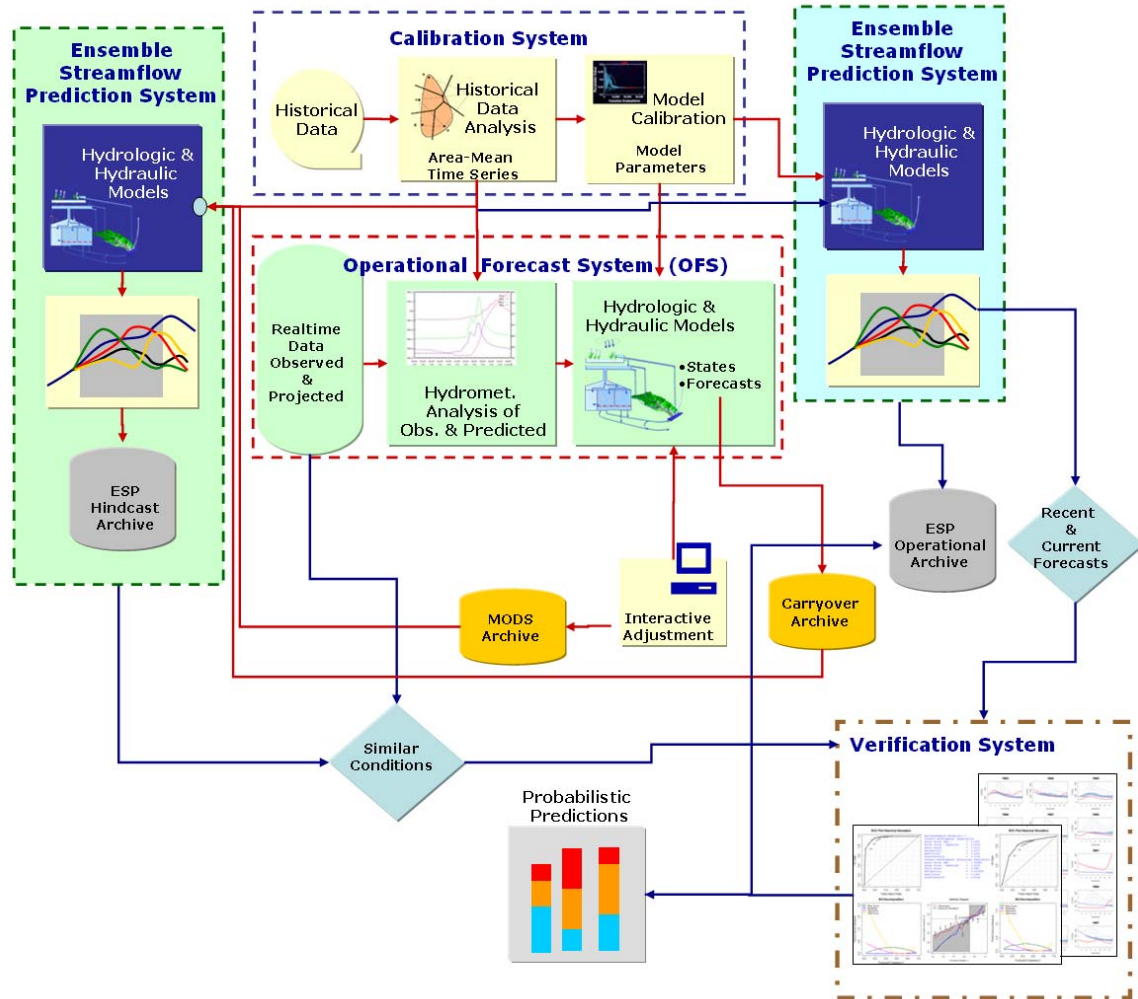


Figure 29. Possible relationships between verification system and real-time/short term operational ESP forecasts. Notice that we not only propose establishing archives of ESP and recent forecasts, but also highlight the need to connect these archives with the ESP generation tools and verification tool.

As seen in the figure, an envisioned (idealized) operational procedure will consist of the following steps

1. In advance, conduct hindcasts (365 days for the period of record). Whenever possible include MODS into the hindcasts, as well as historical data, and /or PQPF. Hindcast procedure should emulate the actual forecasting procedure, whenever possible. Initially, this may have to be conducted manually, or tools that integrate actual carryover, MODS into the ESP procedure, as well as optimized model states.

Hindcasts must be saved not in their probabilistic or graphical formats, but in a format that includes the full ensemble information in their raw format or in flat ascii files (for ease of possible integration with R statistical package)

2. In advance, archive operational ESP, using the same format as above, making sure to archiving full ESP ensemble files rather than derived graphical products or exceedence information.
3. Perform deterministic standard simulations and subsequent ESP runs for the desired duration
4. Query, manually or automatically, dates with similar conditions (procedure needs to be established for criteria of similarities)
5. Query ESP-hindcasts for classical hindcasts as well as those for similar conditions
6. Perform verification, if needed during verification adjust
  - a. Probability distribution (normal, empirical, others)
  - b. Weights of ensemble members (Werner, 2004)
  - c. Forecast probability ranges
  - d. Predictand (the variable being forecasted)
7. Once satisfied with obtaining best verification measures, apply above adjustments to current ESP traces, and issue probabilistic forecast.

---

## **VII. REVIEW OF RIVER FORECAST VERIFICATION DRAFT REPORT**

---

This review was conducted after several discussions with members of the NWS-verification team. We have exchanged drafts of our reports in order to ensure that the work of this team is both consistent and complimentary with the verification strategy adopted by the NWS.

### **VII.1 GENERAL COMMENTS**

Overall the document looks great. It looks like the team really understands the diverse uses of verification, the importance of archiving information, and using a variety of metrics to assess the multi-dimensional nature of forecast performance. There are, however, several additional points to consider.



## VII.2 CRITICAL ISSUES

The Reliability and Discrimination diagrams should be included in the National Baseline Verification System (NBVS). These diagrams are the most informative graphics for evaluating forecast performance. They are being used for climate outlook evaluation by the Office of Climate, Water, and Weather Services (R. Livezey, Climate Services Division, personal communication, 2006), and Franz et. al (2003) demonstrated that these diagrams can verify that hydrologic ESP forecasts contain real information when other metrics can't see it. Experience in conducting training workshops on forecast verification (e.g., Hartmann, 2005, 2006) has established that these concepts are not difficult to understand, and provide a natural lead-in to concepts of Bayesian estimation and adjustment.

Also, the team has it exactly backwards when it comes to comparing deterministic and probabilistic forecasts. Rather than converting probabilistic forecasts to deterministic form (ignoring the essential information provided by the probabilistic forecast), it's more appropriate to convert the deterministic forecasts to probabilistic form. The deterministic forecasts were never without uncertainty; it was just typically not communicated. The deterministic forecasts can, relatively easily, be converted to probabilistic form by overlaying an estimated error distribution around the deterministic forecast value. An easy and reasonable error distribution comes from the calibration statistics. The seasonal water supply outlooks have already been issued in this form for years (with their 90,70,30, and 10% exceedance quantiles). Even though a lot of people are using the median of the probabilistic ensembles, it doesn't mean the approach is correct -- it's not correct. Let's recognize that the deterministic forecasts always did have a probabilistic interpretation and compare them on that basis. Let's not throw away the whole purpose behind probabilistic forecasts, especially considering how much more 'expensive' they are to support.

## VII.3 OTHER ISSUES

On pages 21/22, there's mention of a flexible verification system that would let a user define the statistical variable of the forecast to be verified. Then it mentions, as examples, the mean, median, maximum or minimum value during that time interval. While that's ok, the phrasing suggests that, again, the team may have missed the key contribution of probabilistic forecasts and their essential nature. That is, evaluating a specific value using metrics designed for deterministic values is inappropriate. Instead, the flexibility should be achieved by creating a system that reinforces the proper interpretation and application of the forecasts (i.e., making tradeoffs between forecast confidence and the range of values). We suggest phrasing along the lines of "flexible selection of parts of the forecast or observation distributions, e.g., allowing users to specify probability and variable intervals, or to specify low-flow thresholds".

It would be helpful to stress the benefits of considering the distributions of model inputs and state variables as part of verification (i.e., validation). It would be useful for forecasters in assessing the realism of specific ensemble traces to be able to visualize where the model inputs and state variables lie within their own historical distributions. Those historical

distributions may be based on observations (model inputs) or simulations (internal state variables).

Other visualizations should be considered part of forecast validation as well. First, the concept of visualizing the forecast system status relative to each component's historical distribution should be extended to forecaster run-time MODS, enabling each forecaster to build their own archive that, over time, reflects their tendencies and helps place any real-time MODS choices in perspective with their past choices. Second, the forecasters should be able to visualize the evolution of operational forecasts for a common forecast period, i.e., as the lead-time diminishes.

In the section on the "Review of the available verification tools", Table 3 lists a series of projects and then the supporting text says, "These existing projects will be used to define all the RFC hydrologic verification system requirements." This sounds like there's no place for any activity outside these projects. Is that correct? The report mentions the need for research activities in several places, but it would be useful, from an academic perspective, for the report to explicitly state somewhere (probably in this section) that there is a recognized need for external research and partnership research.

Finally, the report represents an important effort on the part of the Verification Team and the NWS. It's exciting to see that the Hydrology component of the NWS is on track to make a real commitment to verification. Now it will come down to how much of a commitment the NWS can really make, and with how much flexibility and openness to outside participation.

---

## VIII. RECOMMENDATIONS

---

### VIII.1 GENERAL RECOMMENDATIONS

1. Hydrologic verification is vital, viable, and possible. Both probabilistic and deterministic hydrologic forecasts can and must be verified.
2. Hydrologic forecast verification must be conducted in the terms by which the forecasts were created. This requires that researchers become knowledgeable of operational forecasting procedures.
3. Mechanisms should be developed to facilitate bidirectional technology transfer, which includes the transfer of operational forecasting technologies to the research community.
4. A survey of all RFC forecasters should be undertaken to identify field priorities and areas where training is required. A suggested survey is included in this report.

## **NWSRFS-VERIFICATION TOOL**

1. Probabilistic forecast paradigm must have equal weight in operational procedures as deterministic forecasting paradigm.
2. Tools to generate probabilistic forecasts must be integrated into the standard operational forecasting tools (IFP\_Map).
3. Operational ESP verification/validation tools should:
  - a. allow for visual inspection of the data
  - b. Incorporate a hierarchy of forecast verification measures with option for increasing complexity.
  - c. be dynamically linked to both ESP archives, and to ESP generation tools
  - d. incorporate means to conduct conditional verification (e.g., similar initial conditions, weather patterns, snow cover conditions)
4. ESP re-forecasting tools must be linked to MODS, QPF, and operational input archives
5. Hydrologic verification tools must be able to jointly address and compare both probabilistic and deterministic forecasts. However, when comparing performances, probabilistic forecasts should not be converted into deterministic forecasts, rather, the uncertainty bounds around the latter will provide for means to address them in probabilistic form.

## **DATA REQUIREMENTS**

1. Maintenance of ESP forecast archives along with archives of MODS and model states (carryover) files is essential to ensure that verification is carried out in manners consistent with forecast issuance.
2. There is a need to establish alternatives to ESP historical initial condition that better simulates forecasters behavior (see section).
3. Sample size issues must be considered in probabilistic verification studies.