Verification of National Weather Service River Stage Forecasts

Edwin Welles, Neftali Cajina, and Hank Herr
Hydrology Laboratory, Office of Hydrologic Development,
National Weather Service, NOAA
Silver Spring, MD

Abstract: The National Weather Service (NWS) has initiated an effort to verify the river stage forecasts issued by the NWS River Forecast Centers. This effort has brought to light the difficulties associated with evaluating river forecasts at single points and aggregating scores over multiple points. As a first step toward developing a coherent set of statistics that effectively capture the quality of a forecast sample set, the NWS has started collecting data and evaluating statistics to be used to analyze the data. Simple statistics such as the maximum error, root mean squared error, and average error are considered, while the more traditional meteorological measures, such as probability of detection and false alarm rate, are also calculated. Furthermore, more sophisticated distributions based approaches and transformations such as the normal quantile transform are also investigated. A demonstration of a software package designed to enable forecasters to answer some of these questions will be provided.

INTRODUCTION

Verifying River Forecasts at the National Weather Service (NWS): Verification is an essential part of forecasting. Analyzing how well forecasts compare to observed values will improve forecasts and will provide a better idea of where to focus resources. Analyzing verification results can shed light on what type of environments are more difficult to model and therefore, need further study. Unlike verification of meteorological forecasts, there are no well-defined procedures for verifying river forecasts. Recently, the NWS River Forecasting Centers (RFCs) began collecting data and using software developed at the RFCs and the Office of Hydrologic Development (OHD) for river forecast verification. One component of OHD developed software that incorporates some of the verification procedures in meteorology is the Interactive Verification Program (IVP). The IVP computes different statistical measures and displays the results in a variety of graphical formats. It allows an RFC forecaster to make objective evaluations of their forecasts and share their results with other RFCs and the public. One of the goals of the NWS Hydrologic Verification Program is to determine which statistics are best suited for aggregation across numerous forecast points with the goal of being able to characterize the hydrologic forecast skill on a national basis. Although the current system can compute statistics for any number of forecast points and forecast-observation pairs, it is not clear which statistics are best suited for the characterization of forecast quality on a national scale.

INTERACTIVE VERIFICATION PROGRAM

Analyzing Verification Results in the IVP: Analyzing forecast-observation pairs in the IVP begins with a simple scatter plot. The data for each forecast point are displayed as a different color and symbol so outliers can be identified. The scatter plot provides an initial comparison of forecast point quality. It also helps forecasters identify forecast points with obvious biases possibly related to model parameters or inputs. After visually inspecting the scatter plot, the user determines which of the forecast points to analyze in greater detail.

In the next step in the analysis the users select regions for data stratification. Each region is a conditional distribution as described in Murphy and Winkler (1987). The conditional distributions are based on p(f | o) or p(o | f) where o and f are the observations and forecasts depending on whether the user decides to create intervals using the range of observed or forecasted values. For each region, a set of statistics is computed and displayed. Figure 2 shows some of the more traditional verification measures (Root Mean Square Error, Mean Absolute Error, Mean Error, and Maximum Error) for each category.

Figure 3 displays some other measures commonly used in meteorological verification (Probability of Detection (POD), Traditional False Alarm Rate (TFAR), and Average Lead Time of Detection) also based on the regions the user has selected, as well as some hydrological adaptations of these scores (Hydrologic False Alarm Rate (HFAR), Under Forecast Rate (UFR), Over Forecast Rate (OFR), and the Average Lead Time of Detection). Computing these measures consists in arranging observation and forecast data into a three category contingency table (Table 1), and then aggregating the data to calculate different percentages.

- The POD for a region is computed as the percentage of times when the forecast was for that category and the observed is also in that category.

  [E/(B+E+H)]

- The TFAR for a region is computed as the percentage of times a forecast is within an interval but the observation is not.

  [(D+F)/(D+E+F)]

- The HFAR is an adaptation of the TFAR to make it more hydrologically relevant. It is a measure of the number of times a forecast falls below a particular category and it is computed as

  [D/(D+E)]

- The UFR is the number of times the observed is within the interval and the forecast is below the interval, divided by the number of times the observed is within the interval.

  H/(B + E + H).

- The OFR is the number of times the observed is within the interval and the forecast is above the interval, divided by the number of times the observed is within the interval.

  B/(B + E + H).

- The Average Lead Time of Detection is computed as the average lead time of all forecasts that fall into the correct observed category.

	Observed was...		
Forecast was...	Below	Within	above
above	**A**	**B**	**C**
within	**D**	**E**	**F**
below	**G**	**H**	**I**

Table 1: Three Category Contingency Table

There are also distributions based measures the forecaster can use to take a further look at the conditional distributions of the forecasts-observation joint distribution. This distributions oriented approach follows the work of Murphy and Winkler (1987). The focus in this sort of evaluation is on characterizing the distinctness of each conditional distribution. If the conditional distributions are similar, then the forecasts have little skill, while conditional distributions that are dissimilar indicate skill. The degree of similarity and the variation in the conditional distribution characteristics can be used to assess the overall forecast skill.

The first plot in this series is a box and whisker plot. It provides a summary of the conditional distributions and by inspection one can determine if the conditional distributions tend to lie one atop the other, or if they appear distinct. The second shows the actual conditional distributions, which provide the most complete look at the forecast-observation relation. However to the uninitiated, these plots are barely readable. Recently, computation of the Kolmogorov-Smirnov two-sample test statistic was added to IVP. This distributions based measure determines whether the observation and forecast distributions are statistically different for a pre-defined level of significance.

EXAMPLE

The following example will illustrate the process. Consider a set of forecasts and observations with lead times up to three days over a three month period. The forecasts are from forecast points on different rivers which have different flow regimes and flood heights. The samples for each forecast point appear to be from different distributions making it difficult to aggregate verification scores from single forecast points into regional values. The Normal Quantile Transform (NQT) (Kelly and Krysztotowicz, 1997) has been investigated as a means of normalizing the data, but the process is limited because there is insufficient climatological data. Other transforms are possible; for example rating curves can be used to move the stages into flows. However, changes to the rating curves make this transformation difficult to implement. With all normalizations, it is difficult to transform the data back to stage because it is not clear how to transform error measures in a normalized space back to stage because the computed statistic is a derived value. It may be necessary to provide users with unit independent measures, but in general it is desirable to provide skill measures in the units of the forecast.

In order to simplify the selection of points to be analyzed from the scatter plot, the forecast points are identified by color and symbol as is shown in Figure 1. A single point from the Alaska River Forecast Center is selected. It has aggregate statistics of RMSE = 0.8, Mean Error

= -0.1. In many instances, the RMSE and Mean Error are used to define the quality of forecasts for a research simulation. As will be demonstrated these two statistics are not sufficiently informative to characterize of the skill of the forecasts at this location.

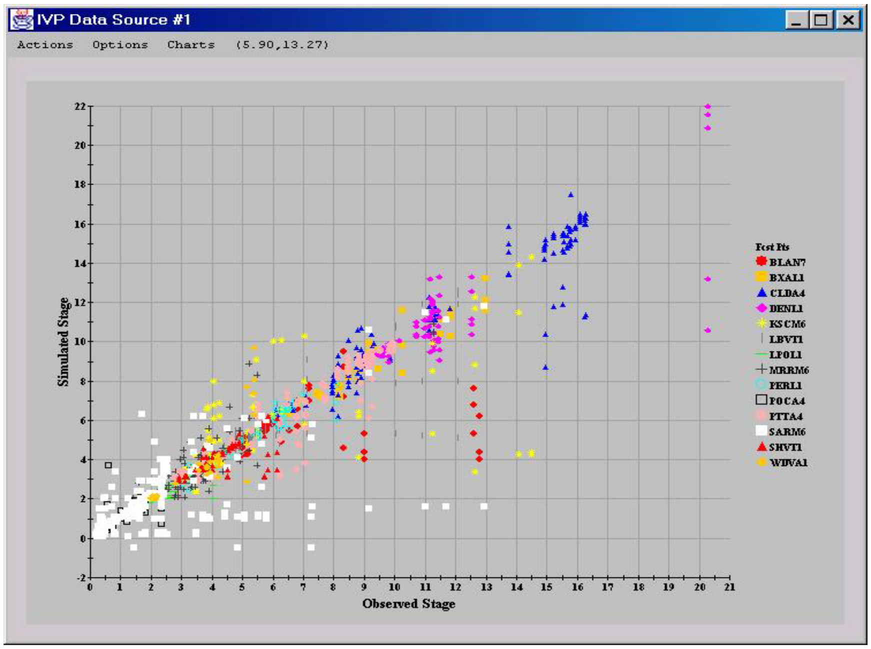

Figure 1: Scatter plot for multiple forecast points

The importance of looking at categories within the total sample of forecasts becomes very clear when the forecasts for this sample are divided into 7 equal regions along the observed axis (these are the distributions p(f | o)) with sample sizes varying from 500 at the low end of the forecast range to 30 sample points at the high end. Figure 2 shows the traditional statistics computed for each range.

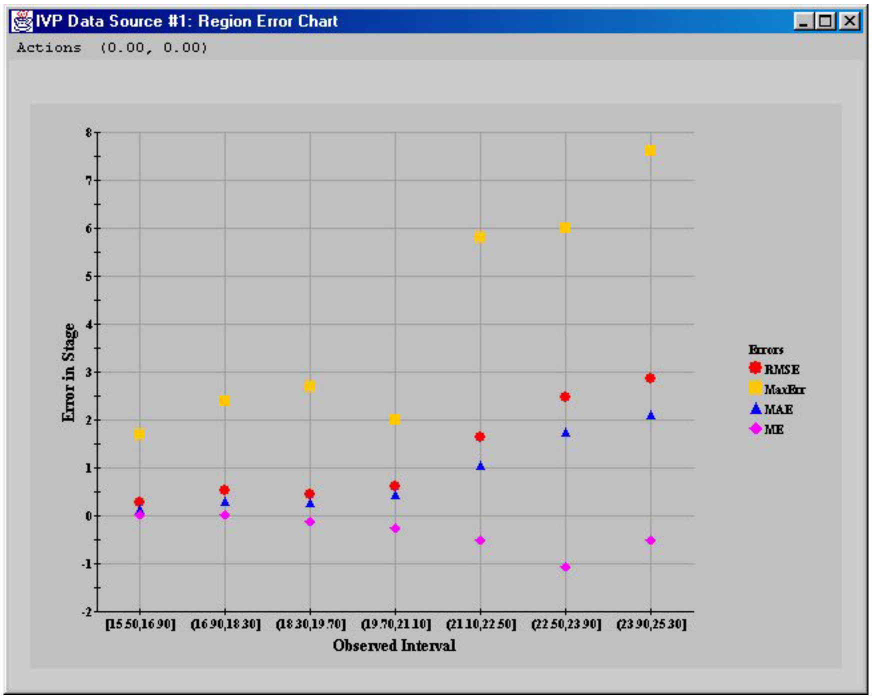

Figure 2: Error by region for one forecast point

The tendency to under-forecast the larger events becomes clear from this simple categorization. This tendency is not captured in the large aggregate and it is the sort of information that can help forecasters focus their efforts to improve the forecast process. It is necessary to break the sample into smaller distributions in order to see the important higher forecasts.

Another traditional set of statistics is the set of the categorical statistics such as POD and FAR. Figure 3 shows the PODs and FARs for the categories. These statistics also indicate that forecast skill decreases with the height of the forecast. The Average Lead Time of Detection is very low at the highest category. The POD decreases and the FAR increases as the observed flow increases. However, these statistics exhibit much more noise than the previous measures. In other words, the trends are less clearly identifiable.

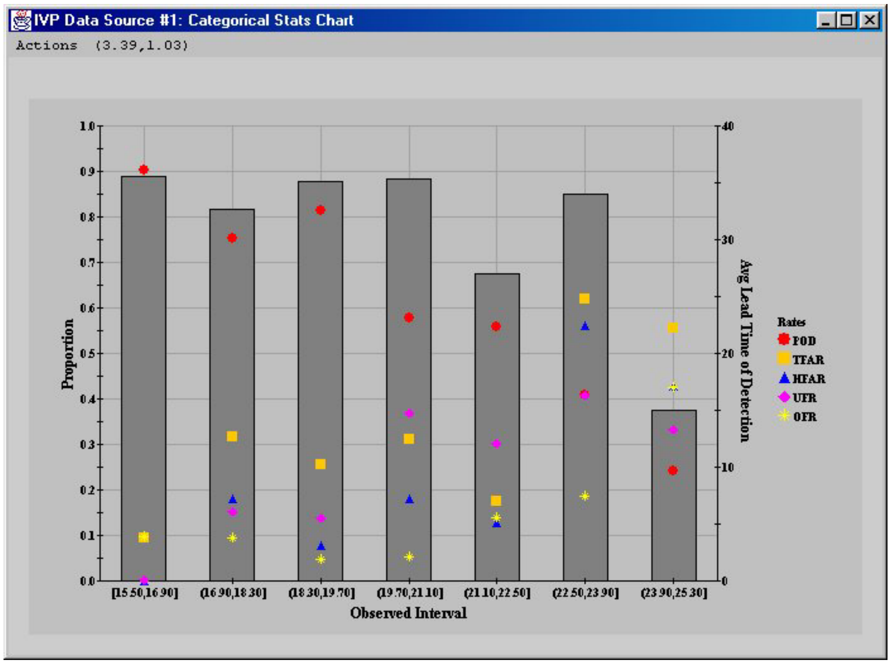

Figure 3: Categorical statistics for one forecast point

The next two figures show the conditional distributions themselves and statistics describing those conditional distributions. In Figure 4 the over forecasting at high stages becomes apparent at this scale as well.

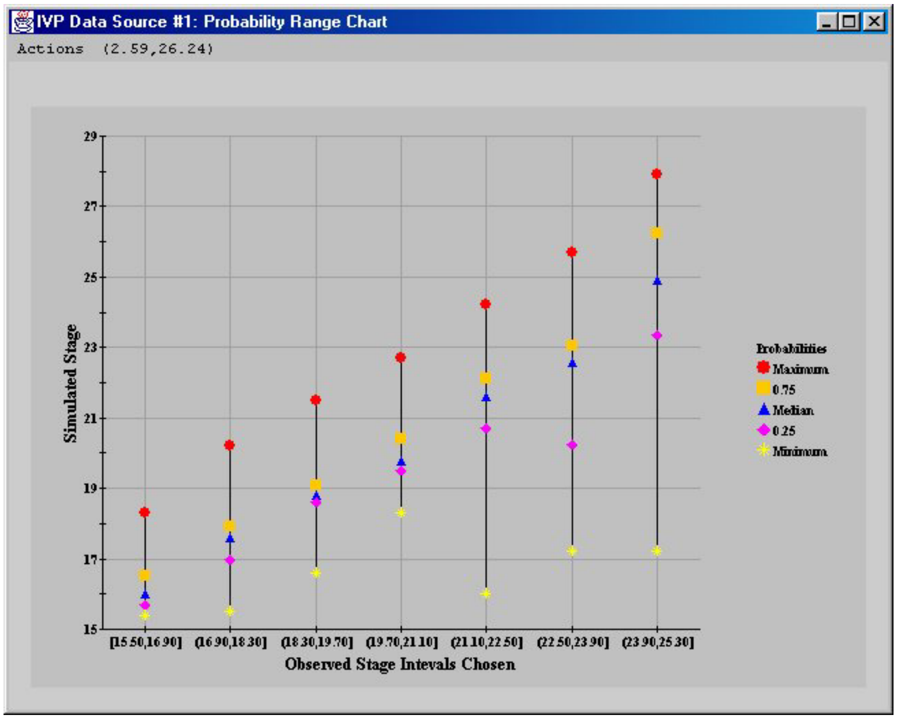

Figure 4: Characteristics for the conditional distributions for one point

More detail about the distributions is provided in Figure 5. The extent to which these plots provide additional information beyond the original error statistics plot in Figure 2 is not clear. However these statistics may provide a method for deriving a more meaningful statistic for large scale aggregations than the standard Mean Error statistics.

Again, it is important to emphasize that one purpose of the Interactive Verification Program is to allow forecasters and others to determine which methods of forecast evaluation work the most effectively. For that reason, many graphics have been provided even if they appear redundant.

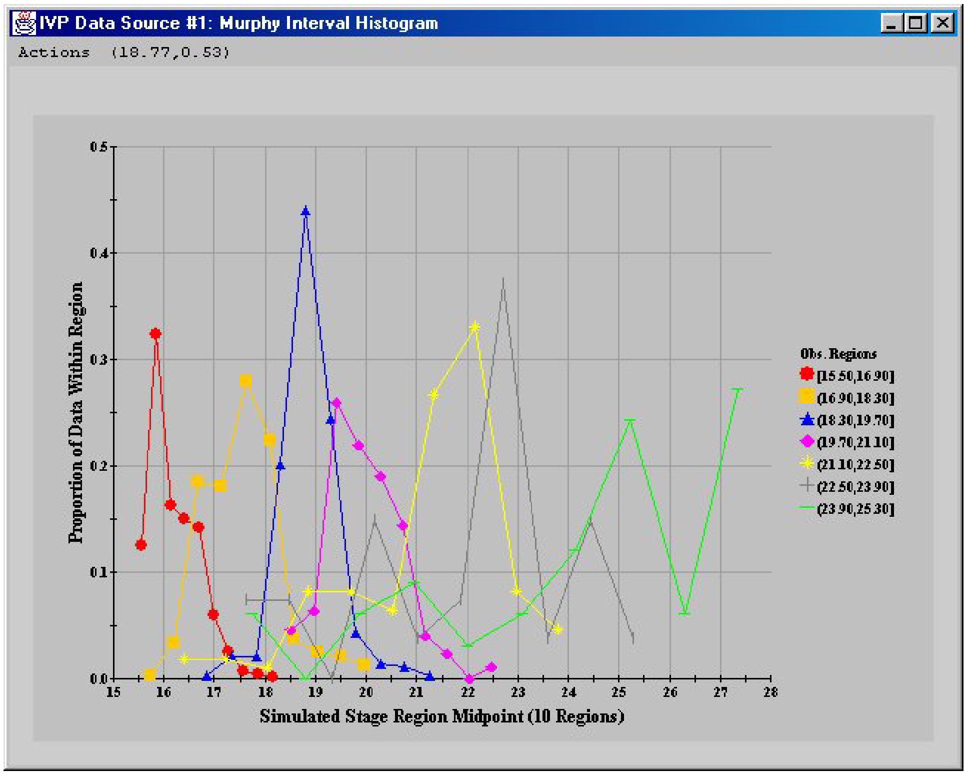

Figure 5: Conditional distributions for one point

ONGOING WORK

The DIS Statistic: Other distributions based scores are also being investigated. Murphy et al. (1989) suggest a score they call a discrimination score, DIS. This score is designed to summarize the distinctness of each conditional distribution. The better the forecasts the more distinct the distributions. The computations of this score have been explored and there appears to be a discontinuity when the forecasts are perfect. Although this is not a likely situation, further work is being done to demonstrate the score behaves as expected.

The No Forecast Problem: Additional work is also required to account for those times when a flood occurs and no forecast has been issued. There are numerous "flood only" forecast points across the United States. These are points for which forecasts are issued only when there is a threat of flooding. In the case where the forecast office does not perceive the threat of flooding, they do not issue a forecast, but a flood may occur all the same. At this time, the verification process depends upon the assumption of a forecast – observation pair.

Timing Errors: It is possible to have the appearance of poor forecasting because the hydrograph is shifted by some number of hours. The peak may be correct, but the time of the peak is off by several hours. In the case of a steeply rising hydrograph, this error, leads to large apparent errors when evaluated from stage perspective. However, from the perspective of time, the errors are not large. The work of Morris (1998) suggests using crossing times to compute timing errors.

Rising and Falling Limbs of the Hydrograph: A final very important area of work required to make the verification process more robust is sorting the forecasts into the rising and falling limbs of flood hydrographs. Models perform markedly differently in the rising and falling limbs of hyrdrographs, and it is important to sort between the two to effectively evaluate model performance.

CONCLUSION

The National Weather Service has embarked on a national Verification Program for the hydrologic services. This is an important step and will lead to a better understanding of the forecast process and help elucidate approaches to improving the forecasts. In support of this program the OHD Hydrology Laboratory has developed a method that provides an increasingly detailed look at the forecasts and observations. The process has been coded into a Java program called the Interactive Verification Program. One goal of the overall verification program is to provide forecasters with the opportunity to evaluate methods of forecast verification and to permit them the opportunity to develop and test statistics to be used on a national basis.

REFERENCES

Kelly, K.S. and R. Krysztotowicz, 1997. “A Bivariate Meta-Gaussian Density for Use in Hydrology”, Stochastic Hydrology and Hydraulics, 11, 17-31.

Murphy, A. and B. Brown and Y. Chen, 1989. “ Diagnostic Verification of Temperature Forecasts”, Weather and Forecasting, Vol 4, pp 485 – 501.

Murphy, A. and R. Winkler, 1987. “A General Framework for Forecast Verification,” Monthly Weather Review, Vol 115, pp 1330 – 1338.

Morris, D., 1988. “A Categorical Event Oriented Flood Forecast Verification System for National Weather Service Hydrology,” NOAA Technical Memorandum, NWS Hydro 43.