463 VERIFICATION OF ENSEMBLES AT THE MIDDLE ATLANTIC RIVER FORECAST CENTER

Andrew W. Philpott*, Patricia Wnek, and James D. Brown
NOAA/NWS, Middle Atlantic River Forecast Center, State College, PA
NOAA/NWS, Silver Spring, MD and University Corporation for Atmospheric Research

1. INTRODUCTION

Four National Weather Service (NWS) River Forecast Centers (RFCs) are producing short-term (0-7 day) streamflow forecasts using meteorological model ensembles. Middle Atlantic, Northeast, Southeast, and Ohio RFCs will continue to generate products in an experimental period lasting through September 15, 2012, during which time public comments will be solicited.

The ensemble forecast products are available at http://www.erh.noaa.gov/mmefs/.

Three different ensemble forecasts are available, based on three different National Centers for Environmental Prediction (NCEP) meteorological ensembles, namely the 21-member Global Ensemble Forecast System (GEFS) through lead time 168 hours, the 21-member Short Range Ensemble Forecast (SREF) system through lead time 90 hours, and the 42-member North American Ensemble Forecast System (NAEFS) through lead time 168 hours. Precipitation and temperature ensembles from these systems are interpolated into basin averaged values and run through the RFC's hydrologic model to produce ensemble forecasts of snow water equivalent and river stage. Plots of the input ensembles and output ensembles are provided at all forecast points on the Ensemble River Forecast webpage.

The focus of this paper is on verification of the streamflow forecasts, however work is also ongoing in verifying basin averaged ensemble precipitation and temperature inputs. The Middle Atlantic RFC created four-times-daily SREF based ensembles from December 2008 through November 2011, which included about 4200 forecasts. NAEFS and GEFS based ensemble river forecasts are only twice-daily, and were only available for one year, from November 2010 through November 2011. Since ensemble forecast verification relies on a large sample size, it was most productive to focus this paper on the SREF ensembles. However, limited comparisons were made between the SREF, NAEFS, and GEFS ensembles over the one year period in which all three were available.

2. METHODS

All verification work was completed using the Ensemble Verification System (EVS) developed by the NWS Office of Hydrologic Development in Silver Spring, MD (Brown et al. 2010).

The first step in verification pairs forecasts with observations. A variety of river forecast points were selected, including larger rivers and smaller headwater locations. Temperature and precipitation verification required basin-averaged precipitation and temperature data from the Middle Atlantic RFC archive to be paired with the basin-averaged ensemble input forecasts. Six-hourly river stage observations from the USGS and NWS were converted into streamflow (in meters cubed per second) using archived USGS rating curves, to be paired with the streamflow ensemble forecasts.

The coefficient of variation of the root mean square error, relative mean error, and correlation coefficient of the ensemble mean forecast were computed with the EVS. Coefficient of variation of the root mean square error and relative mean error are both normalized metrics such that they can be compared between locations having greatly different magnitudes of streamflow.

Box plots of the ensemble forecast error against observed values were made to investigate spread and bias. In addition, spread-bias diagrams displayed how spread and bias impacted forecast reliability, by plotting the frequency at which observations fell within each portion of the ensemble distributions. Biases or underspreading would cause the observations to fall more often within a certain "window" of the ensemble distribution than implied by the size of that window, for example being above the 75th percentile more than 25% of the time.

Forecast reliability and discrimination are two of the key components to determine ensemble forecast quality (Wilks 2006). Forecast reliability indicates how well the forecast probabilities match up with the observed relative frequencies on those occasions when something is forecast with a given probability. The reliability diagrams compared the forecast probabilities of a particular event to the observed relative frequencies, where the event in this study was defined as the exceedence of the streamflow threshold corresponding to the 95th percentile of the climatological distribution (Hsu and Murphy, 1986). Forecast discrimination indicates how well-differentiated the forecasts might be given different observed conditions. For example, the Relative Operating Characteristic (ROC) compares the probability of detection, which is the fraction of observed events that were correctly forecast to occur, against the probability of false detection, which is the fraction of observed non-events that were incorrectly forecast to occur.

* *Corresponding author address:* Andrew Philpott, NOAA, National Weather Service, Middle Atlantic River Forecast Center, 328 Innovation Blvd, Suite 330, State College, PA, 16803;
email: andrew.philpott@noaa.gov

3. RESULTS

a) Comparison between GEFS, NAEFS, and SREF

Over the one year in which GEFS, NAEFS, and SREF ensembles were all available, they were similar, except that the SREF ensembles had a larger negative bias early in the forecast period (Fig. 1). This problem with the SREF forecasts may be related to how the ensemble river forecasts handle SREF runs that start three-hours off the standard river forecasting center synoptic times of 0Z, 6Z, 12Z, and 18Z (which appears to introduce an underforecasting error in the precipitation forecasts on the first timestep). The exact cause and a solution to this bias in the ensemble river forecasts must be determined. By day 3, the three different ensemble types had very similar relative mean error.

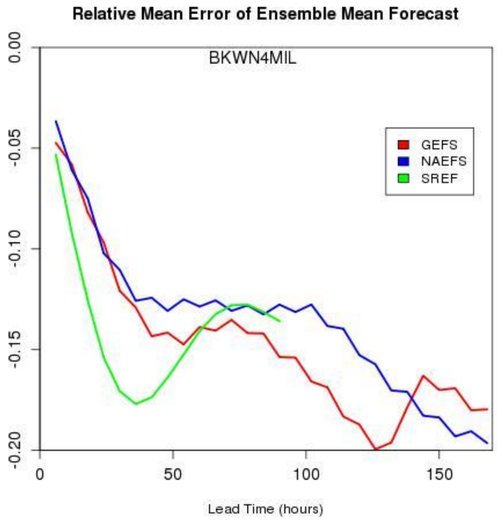


Figure 1: Relative mean error in the streamflow ensemble mean forecasts for November 2010 through November 2011 at the Millstone River at Blackwells Mills, NJ.

Further comparison of the ensemble river forecasts reveal that the GEFS-based forecasts had less spread in the forecasts than the NAEFS- and SREF-based forecasts. This is expected, since the SREF and NAEFS ensembles are based on multiple meteorological models. The NAEFS- and SREF-based streamflow forecasts, but particularly the GEFS-based forecasts, are underspread which affects the reliability of the ensemble streamflow forecasts.

b) Ensemble Mean Statistics in SREF over 3 years

SREF ensemble mean forecasts from December 2008 through November 2011 were compared between different locations (Figs. 2-4). The relative mean error in the ensemble mean forecasts at five different river forecast points are presented (Fig. 2). Some of the differences between points were due to differences in the bias of the input precipitation forecasts. However, there were also differences in the hydrologic modeling biases of different locations that have an effect. For example, Cortland, NY (CRTN6TGH) was the most positively biased of the five streamflow forecasts (Fig. 2), but this was not explained by a difference in bias in the precipitation (not shown). However, Farmville, VA (FARV2AMX) was the most negatively biased of the locations both in precipitation (not shown) and streamflow.

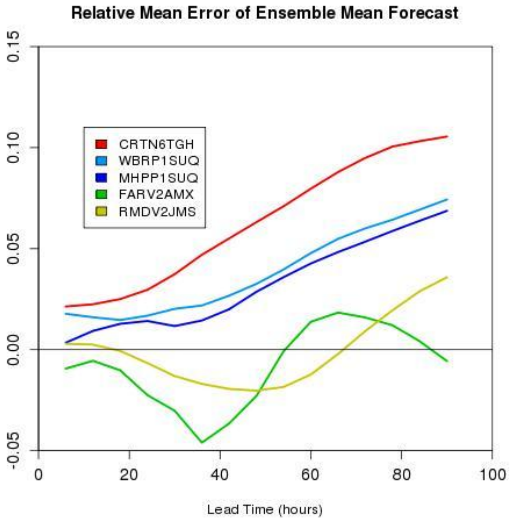


Figure 2: Relative mean error in the SREF-based streamflow ensemble mean for November 2008 through November 2011 at five river forecast points. Locations key:
CRTN6TGH: Tioughnioga River at Cortland, NY
WBRP1SUQ: Susquehanna River at Wilkes-Barre, PA
MHPP1SUQ: Susquehanna River at Meshoppen, PA
FARV2AMX: Appomattox River at Farmville, VA
RMDV2JMS: James River at Richmond, VA

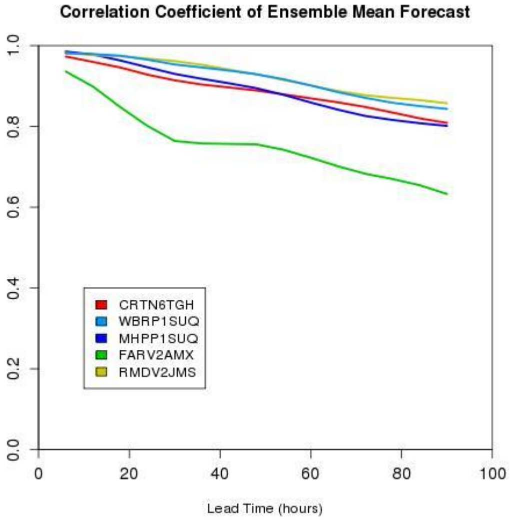


Figure 3: Correlation between SREF-based streamflow ensemble mean forecasts and observations.

The linear correlation between observations and the ensemble mean forecasts was high (Fig. 3), although it decreased steadily with forecast lead time. Curiously, FARV2AMX (Farmville, VA) had a lower linear correlation between forecasts and observations than the other four points. One possible explanation is greater hydrologic modeling errors at this location, perhaps related to hydrologic model parameters. However, particular problems with the precipitation forecasts at this location, and with the basin-averaging process (the Appomattox River basin has a relative scarcity of rain gages compared to other Middle Atlantic RFC basins), also may have contributed.

Finally, the Root Mean Square Error (RMSE) shows the average error-spread of the ensemble mean forecast (Fig. 4). The subsequent graphs of RMSE have been divided by the mean observed flow at each point over the 3 years to allow for comparing between locations (i.e. the coefficient of variation of the RMSE). The river forecasts at Farmville, VA had a higher coefficient of variation of RMSE than the other location, similarly to how they had a lower correlation coefficient. At all points, there was a steady degradation in forecast quality with increasing forecast lead time, due to a reduction in the quality of the SREF forcing forecasts for longer lead times as well as cumulative hydrologic model and basin averaging errors.

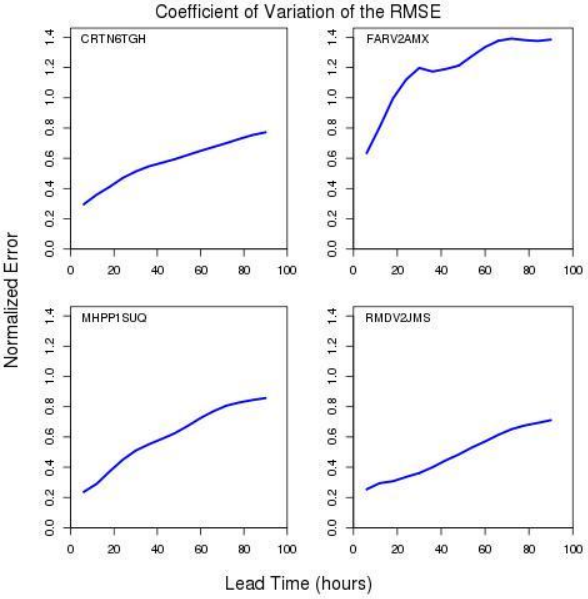


Figure 4 Coefficient of variation of the RMSE of the SREF-based ensemble mean forecasts at 4 river forecast points.

c) Biases and Spread

Box plots of ensemble forecast error versus observed values are capable of showing biases, both unconditional and conditioned on the observed value. They also give an impression of any problems with overspreading or underspreading of the forecasts. In these plots, each box is a single ensemble forecast of 21 members that was paired with a particular observation (Figs 5-7). The lowest member will have the most negative or least positive error compared to the observation and the highest member will have the most positive or least negative error. The perfect forecasting system would have all boxes intersecting the zero error line. Also, in looking at all the boxes overall, the zero error line would fall within each portion of the various boxes an equal percentage of the time. For example, the median forecast, as indicated by the black dot on the plots, would be an underforecast as often as an overforecast. Furthermore, to have no conditional bias, the chance that the zero error line would fall within any portion of the box would not depend on the observed value at all, so for example there would be just as much a chance for the median forecast to be above the zero error line for high observations as for low observations.

In this case, the forecasts had a conditional bias on the observations, where highest observed events tended to be underforecast by the ensemble median, and even by higher ensemble members. Lower observed events tended to be overforecast by the ensemble median and often by the lower ensemble members as well. Part of the problem can be traced to a strong conditional bias in the precipitation forecasts (Fig. 6). Basin-average precipitation forecasts interpolated from the SREF grid were unable to capture the small scale features that lead to high observed basin-averaged precipitation amounts. However, at this location (Cortland, NY), the hydrologic modeling system introduced positive bias, as seen in the ensemble mean forecast for this location, with the result that underforecasting bias was only present for the very highest observed streamflow cases, while an overforecasting bias was present both for moderate and low flows.

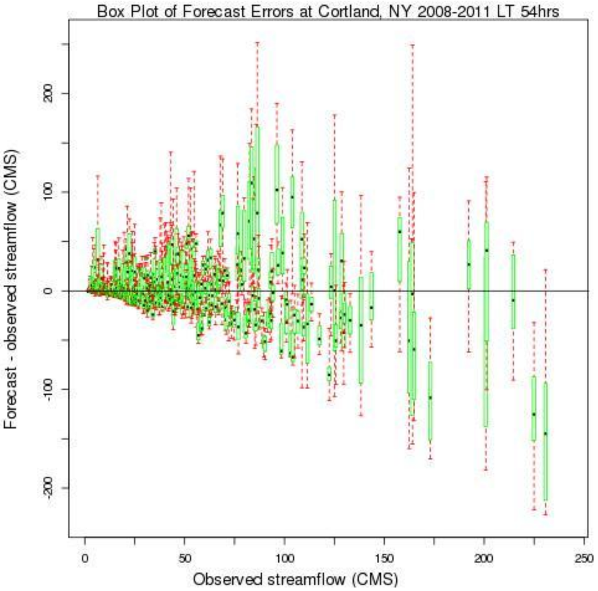


Figure 5: Box plots of the SREF-based input streamflow ensemble forecasts for November 2008 through November 2011 at the Tioughnioga River at Cortland, NY, at a lead time of 54 hours. Each box is a single ensemble forecast of 21 members, which pairs with a particular observation. The green boxes indicate the 25th through 75th percentile of the ensemble distributions and the black dots are the median forecasts. The red whiskers indicate the minimum and maximum forecasts.

The streamflow forecasts also showed evidence of underspreading, with the spread of the forecasts not being large enough to capture the observations in too many cases. This means that the ensemble forecasts were overconfident, in that they sometimes forecast a 100 percent chance that the observation would fall within a range that ended up being not wide enough. The underspreading was worse for the early lead time forecasts (not shown), although the magnitude of errors in early lead time forecasts was less. The cause of underspreading was that the river ensemble forecasts did not account for the hydrologic uncertainties, including those in the initial conditions of the hydrologic model, only the forcing uncertainties. By later lead times, spread in the precipitation forecasts had increased the spread in the streamflow forecasts, decreasing the overconfidence.

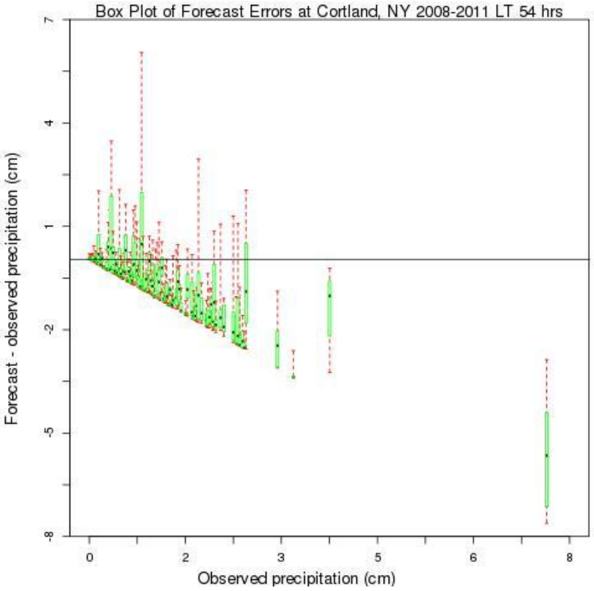


Figure 6: Box plots of the SREF-based input precipitation ensemble forecasts from November 2008 through November 2011 for the Tioughnioga River at Cortland basin, at a lead time of 54 hours. The green boxes indicate the 25th through 75th percentile of the ensemble distributions and the black dots are the median forecasts. The red whiskers indicate the minimum and maximum forecasts.

Overall one of the main conclusions of verification work with the ensemble river forecasts so far, is that a spread based entirely on uncertainty in the precipitation and temperature forecasts cannot completely cover the total uncertainty of river forecasting. Uncertainties in hydrologic factors such as routing, runoff efficiency, effects of rainfall intensity on runoff, snowmelt parameters, and baseflow recession are considerable, and, therefore, the ensemble river forecasting system will be underspread and will have biases specific to each forecast point. Due to hydrologic modeling uncertainties, these ensemble river forecasts cannot capture the full range of possible flows in a particular situation.

Some qualitative comparison of the verification of different locations can be made by comparing the box plots. In this example, there were some differences evident between Cortland, NY (Fig. 5) and a further downstream point at Meshoppen, PA (Fig. 7). The most obvious result was a greater underforecasting bias of the highest observed events at Meshoppen. There also appeared to be somewhat greater underspreading in the Meshoppen forecasts (Fig. 7).

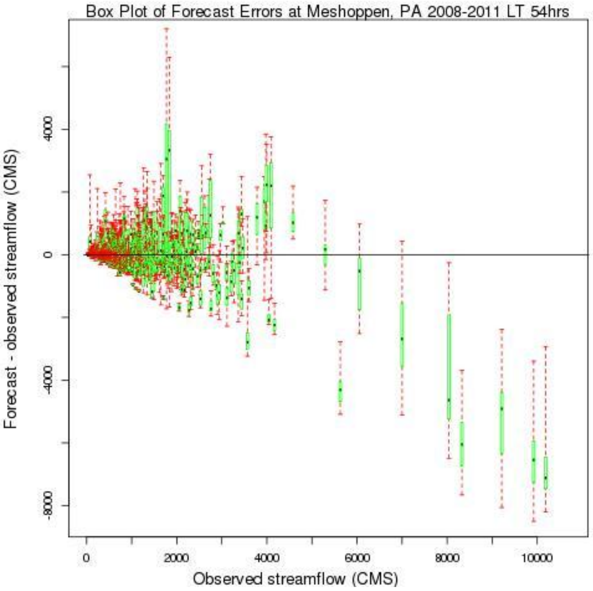


Figure 7: Box plots of the SREF-based input precipitation ensemble forecasts for November 2008 through November 2011 at Meshoppen, PA, at a lead time of 54 hours.

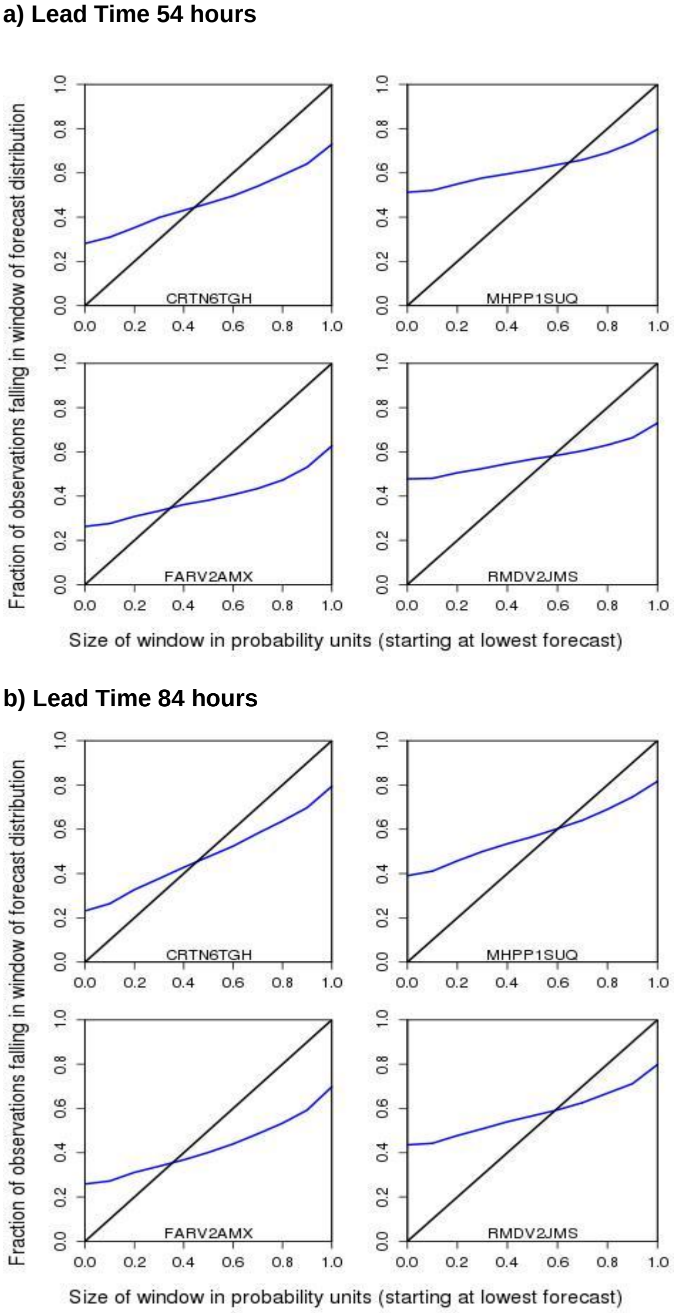
As further evidence of the problems with forecast spread, we conclude this section with a presentation of the spread-bias diagrams, which show how spread and bias in the forecasts affects reliability (Fig. 8). The spread-bias diagram shows directly the probability of the observation (the zero error line in the box plots) falling within each portion of the forecast ensembles. As discussed, the observation should fall below the median forecast 50 percent of the time. If this happens less than 50 percent of the time, this means that the ensemble median tends to underforecast the observation. Furthermore, the observation should fall below the 10th percentile forecast 10 percent of the time, and below the 90th percentile forecast 90 percent of the time.

This perfectly reliable forecasting system is represented by the y=x line in the spread-bias diagrams. The extent to which the forecasts are unreliable due to spread and bias problems determines how much the forecasts deviate from y=x. In this case, the forecasts were more horizontal than y=x. This is a result of both underspreading and conditional bias, as was seen in the box plots. An overspread forecast system (underconfident) would have a spread-bias diagram with a steeper slope than y=x.

These plots confirm that underspreading was more severe at shorter lead times than at longer lead times (Fig. 8), since the 54 hour lead time spread-bias diagrams have slightly less of a slope than the 84 hour lead time plots. The 6 hour lead time plots (not shown) were almost completely flat, with very few of the observations ever falling within the extremely narrow ensemble spreads. Since all spread within these ensemble river forecasts are due to spread within the precipitation and temperature inputs, it makes sense that underspreading would decrease with lead time due to accumulating increase in spread. Also, the spread-bias diagrams confirm that underspreading was more severe at Meshoppen, PA than at Cortland, NY.

Spread-bias diagrams are capable of showing a conditional reliability of the forecasting system (Brown 2010).

A forecasting system cannot be perfectly reliable unless the observations have an equal chance of falling within every portion of the ensemble distributions. The key conclusion from the spread-bias diagrams is that underspreading and conditional bias affected the reliability of the forecasting system, resulting in overconfident forecasts. Ensemble post-processing to correct biases and increase spread could improve these results.



a) Lead Time 54 hours

b) Lead Time 84 hours

Figure 8: Spread-bias diagrams for SREF-based streamflow forecasts for November 2008 through November 2011 at four locations, at lead times (a) 54 hours and (b) 84 hours.

d) Reliability and Discrimination

Reliability diagrams (Fig. 9) provide a more strict evaluation of reliability than the spread-bias diagrams for a particular flow threshold. In this case, the exceedence of the 95th percentile observed flow was chosen, which had a sample size of around 200 observed cases. An ideal sample size would be even larger, but this indicates why a higher flow threshold such as flood stage could not be meaningfully analyzed.

In the reliability diagrams analyzed, there were three bins, one for forecasts that predicted a high chance of exceedence, one for forecasts that predicted a moderate chance of exceedence, and one for forecasts that predicted a low chance of exceedence. The number of cases in which a low chance was forecast was near 4000, and the sample size of the other two bins was 100-200. The sample sizes were plotted in the upper left hand corner of each diagram (Fig. 9) on a log scale, which emphasizes the differences between the moderate and high chance bins. An average forecast probability was computed for each bin, and these were roughly 90 percent, roughly 50 percent, and roughly 0 percent for the three bins. The reliability diagram compared these forecast probabilities to the average observed probabilities. A forecasting system perfectly reliable at predicting the threshold flow would have forecasted probabilities the same as the observed probabilities, which would yield a reliability diagram that matches y=x.

These reliability diagrams indicated a tendency to overforecast the chance of exceeding the threshold both when a near 90 percent chance is forecast and when a near 50 percent chance is forecast. The accuracy of the near 0 percent chance is very good, since there are thousands of cases in which the flows were low and it was very easy for the SREF to discern that no heavy precipitation events were coming. The handful of cases noted on the box plots where a highwater event occurred but was underforecast were overwhelmed in the reliability diagram by the thousands of correctly forecast low flow events. However, closer inspection revealed that in most cases, when a near 0 chance of exceeding the threshold was forecast, the observed chance was slightly higher. These reliability diagrams overall indicated, as did the spread-bias diagrams, an overconfident forecast system: when the forecast system expected a high chance of exceeding the 95th percentile flow, the actual chance was not as high. However, this conclusion may not be statistically significant due to the sample size, and an analysis of confidence intervals (which is planned beyond this paper) would be necessary to make that determination.

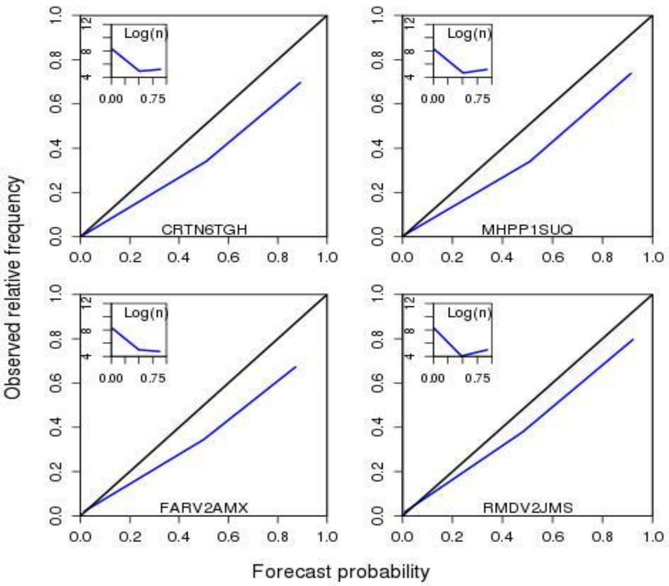
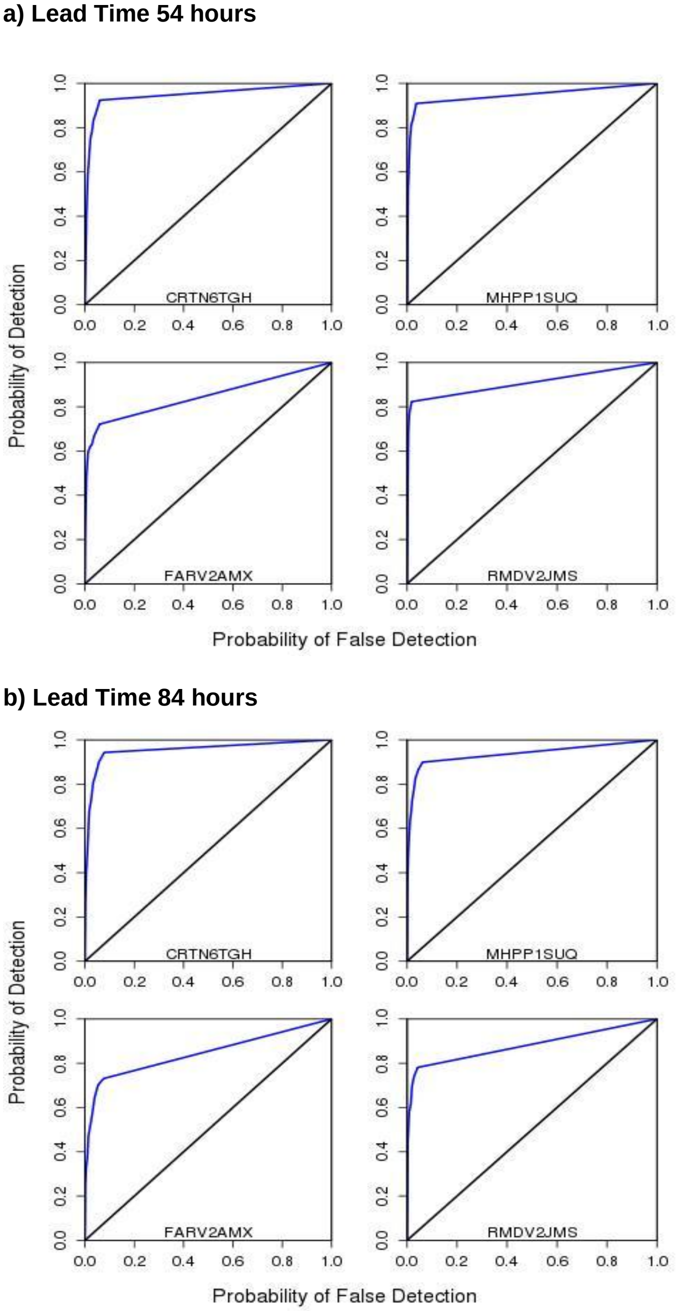


Figure 9: Reliability diagrams for SREF-based streamflow forecasts for November 2008 through November 2011 at four locations, at lead time 84 hours. These diagrams indicate the reliability of forecasting the exceedence of the 95th percentile observed flow at each site. In these plots there are three bins, one with a forecast probability near 0, one near 50 percent and one around 90 percent. The plot in the upper left corner of each diagram indicates the sample size of each bin, on a log scale. The y=x line indicates a perfectly reliable forecast, in which forecasted probabilities and observed probabilities are the same.

Event discrimination was investigated using ROC plots (Fig. 10). These ROC plots were for the same 95th percentile observed flow threshold that was used in the reliability diagrams. They indicated the probability of detection (POD) compared to the probability of false detection (POFD) of exceeding this threshold flow. This was repeated for various "decision thresholds" or binary classifiers at which the forecast probability is interpreted as predicting that the threshold will be exceeded (yes/no). In this context, there is a trade-off between correctly detecting occurrences (having a low detection threshold) and falsely detecting non-occurrences. The line y=x in this case indicates the "random guess" or climatological frequency, where the POD and POFD are the same. A perfect ROC curve would have all points in the upper left corner, at a POD value of 1.0 and POFD of 0.0. In this case, the ROC curves were near the upper left corner, indicating that the forecasting system had a strong ability to discriminate between the occurrence and non-occurrence of the threshold event. Thus, despite the lack of perfect reliability, driven by the overconfidence and conditional biases, these SREF forecasts are considerably more skillful (discriminatory) than climatology (Fig. 10).

The probability of detection at higher decision thresholds appeared to differ between forecast locations. At Farmville and Richmond, the highest POD was only around 0.75, whereas at Cortland and Meshoppen the highest POD was around 0.9. Another interesting result was that the event discrimination was high for a range of forecast lead times, and did not appear to decrease substantially between lead time 54 hours and 84 hours. The statistical significance of these conclusions through confidence interval analysis will be investigated in the ROC plots as well in future work.


a) Lead Time 54 hours

b) Lead Time 84 hours

Figure 10: ROC plots for the 95th percentile flow threshold of SREF-based streamflow forecasts for November 2008 through November 2011 at four locations, at lead times (a) 54 hours and (b) 84 hours.

3. CONCLUSIONS

Verification of streamflow forecasts from the ensemble river forecast system has been completed at several locations in the Middle Atlantic RFC. Concentration was on the SREF-based ensemble forecasts, since four forecasts per day were archived for a three year period, November 2008 through December 2011. There was only one year of GEFS and NAEFS based forecasts. However, comparisons of the SREF, GEFS and NAEFS for this one year period yielded fairly similar results. There was an issue with the SREF-based forecasts on days 1 and 2, which is likely related to the 15Z, 21Z, 3Z, and 9Z start times (when compared to the 0Z 6Z,12Z, and 18Z times of the Middle Atlantic RFC operations) which must be investigated further. However, by lead day 3, the SREF based simulations had very similar mean biases to the GEFS and NAEFS based simulations. Out to lead time 168 hours, GEFS and NAEFS had very similar negative biases in the ensemble mean forecast, although the spread of the GEFS based ensembles was less than that of the other two.

Overall, there was considerable difference between forecast points, especially in probabilities of detection and in biases. Much of this difference can be attributed to errors in the streamflow models and the lack of accounting for hydrologic uncertainties. Model biases in baseflow recession, routing, runoff efficiency, snowmelt parameters all cause uncertainties in the river forecasting that are not captured by simply varying the inputs of precipitation and temperature. In addition to causing differences in forecast skill at different points, this hydrologic uncertainty also resulted in biases and underspreading in the streamflow ensemble forecast at all forecast points. One possible method for improving the forecasts would be ensemble post-processing, to correct biases and increase spread appropriately to attempt to artificially account for and correct for some of the hydrologic uncertainties and biases that the hydrologic ensembling method does not address.

Underspreading and conditional biases resulted in some lack of reliability in these forecasts. This means that when these ensembles forecast a particular chance of exceeding a certain river level, the actual chance may be different. This is a problem that should be corrected to improve the forecasts. Discrimination could also be improved, since neither probabilities of detection nor probabilities of false detection were perfect. However, the forecast discrimination is high, since probability of detection is much greater than probability of false detection.

Some of the conclusions of this paper could be strengthened by assessing the statistical significance of the verification results. For example, we cannot say whether the deviation of the reliability diagrams from perfect reliability is significant. EVS has the capability of plotting confidence intervals, using a bootstrap sampling method. This will be explored in future work.

REFERENCES

Brown, J.D., Demargne, J., Seo, D-J and Liu, Y., 2010: The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Modell. Softw.,* **25(7)**, 854-872.

Hsu, W.-R. and Murphy, A.H., 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2**, 285-293.

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd ed., Academic Press, 627pp.