NWS Hydrology Forecast Verification Team
Teleconference Notes
04/22/2008

Agenda

- Slides on the IVP exercise developed by Julie Demargne
- Presentation of the WGRFC verification case study by Greg Waller

Questions and Comments

Julie's presentation

Slide #5 (Plots #5-7): for these pair plots for the 3 lead days, the 6-hr forecasts for the four 6-hr lead times relative to a given lead day are all pooled together to display the forecast-observed pairs. When computing the verification statistics for each lead day, the statistics will be relative to all forecasts for 6-hr, 12-hr, 18-hr and 24-hr lead times.

Slide #6: these are additional plots for the Real QPF Forecasts showing how much the spread in the pair plot varies for the different 6-hr lead times. The spread varies a lot between individual 6-hr lead times, for low events and more especially for high events. Therefore it would be better not to pool together 6-hr forecasts from different lead times when computing verification statistics for these Real QPF Forecasts since the quality of these forecasts varies with 6-hr lead times.

Slide #7 (Plots #8-9): ABRFC added that the flood event on 06/21 was missed probably due to strong QPF under-forecasting, which is common for convective events in May-June. For the second smaller event on 06/25, it was mostly a timing error, the forecast being 12 hour late.
It is actually difficult to guess what MODs were made by just looking at the time series plot.

Slide #8: this is an additional plot to show the whole period of record for the Real QPF Forecasts using just the first 6-hr forecast values. The forecasts are available from 12/01/1996 to 12/31/2000. This plot is useful to analyze the high events in the whole verification period. In this case, there are only 5 flood events in the whole verification period. Therefore the statistics relative to above Flood Stage are not statistically significant. It would be better to compute statistics for the Action Stage or a lower stage threshold to increase the number of high events.

Slide #10 (Plots #12-15): when the sample size is below 20 to 30 events, the statistics are not robust enough and most variations in the numbers are due to sampling uncertainty. That's why the RMSE for perfect forecasts oscillates with lead times.
For the Real QPF Forecasts (blue curve), the results were quite different when looking at the conditions Obs < FS (Flood Stage) and Forecast < FS. For Obs < FS, the Real QPF Forecasts are slightly worse than the Zero QPF Forecasts from 24 hr to 42 hr lead time; for

Fcst < FS, the results are similar for Real QPF Forecasts and Zero QPF Forecasts (with slightly larger RMSE values than for Obs < FS). This difference between the RMSE for the 2 conditions must reflect the cases when the Real QPF leads to a stage forecast above FS, for which the observation is below FS, leading to worse RMSE values than for the Zero QPF stage forecasts. You can see the forecast-observed points relative to the 2 conditions on the pair plots (see slides #4-5), for which the flood stage is plotted as the green line.

Slide #11 (Plots #16-17): the plots have been modified to include only forecasts from 1997 to 2000 since 1996 has forecasts only for December.
For Plot #17, the statistics for June are much higher than for the other months, which show the difficulty to predict for convective events. It must reflect the missed flood event of 06/21/2000. The good results for the Persistence Forecast in August and December are due to dry conditions at this time of the year and for these 4 years of data.

Slide #12 (Plots #18-19): For Plot #18, the forecast and observed distributions look very similar with a slight tendency to over-forecast. In Plot #19, we can see that these results are dominated by the observed flow events below FS. For the observed events above FS, these statistics are relative to only 5 flood events and therefore are not statistically significant, especially when computed for each individual month. To analyze the forecast performance for high events based on more events, one could define lower stage categories (for example relative to action stage) to increase the sample size.

Slide #13 (Plot #20): the HFAR statistic shows that Zero QPF Forecast performs better than the Real QPF forecasts and similarly to the Perfect QPF forecasts. It reflects the tendency to over-forecast and then forecast more floods than what actually occurs when using the Real QPFs. Again, this is relative to a small number of flood events. It would be better to define a lower stage threshold (for example relative to action stage) to compute the statistics on more high events.

Slide #14: These are additional plots similar to Plot #20 for individual 6-hr lead times and also showing the sample sizes for flood events. Again, the sample sizes are too small to get robust results. And the POD and HFAR vary significantly between individual 6-hr lead times, showing that it is better not to pool the forecasts from 6-hr to 24-hr lead times together when analyzing the forecast quality.

Slide #15 (Plots #21-22): the Perfect Forecasts show almost perfect event discrimination for all lead days. For Day 1, Zero and Real QPF forecasts show similar skill for event discrimination, which is better than for persistence. For Day 3, the Real QPF Forecasts perform better than Zero QPF Forecasts, which is more similar to the Persistence forecast results.

Slide #16: these are some issues and recommendations regarding this verification exercise, which could be useful for other verification case studies.
The number of flood events for this 4-year verification period is too limited to compute significant verification statistics for flood events. In order to increase the sample size when

analyzing the forecast quality for high events, one could define a lower threshold value (such as the action stage), extend the period of record, and/or pool forecast-observed pairs from similar forecast points (although this could be difficult to find similar forecast points, especially for stage of flow).
Also verification statistics relative to a specific lead day are computed by IVP by pooling forecasts from different lead times: for example, the day-1 statistics are relative to all forecasts for 6-hr, 12-hr, 18-hr and 24-hr lead times. However forecast quality varies a lot between individual 6-hr lead times as shown on the pair plots on slide #6 and on the POD-HFAR plots on slide #14. Therefore it is better to compute verification statistics for individual lead times and analyze how the results vary with lead time.

For the RFCs, please send any comments about this IVP exercise to Julie D. (Julie.Demargne@noaa.gov) so that this exercise could be improved.

WGRFC Case Study

There were some comments about how to expand this verification study using the 23 forecast points and analyzing the results according to lead time. Also, in order to compare the VAR forecasts issued every 1 hr and the NWSRFS forecasts issued every 12 hr, the VAR forecasts could be sub-selected to use only the forecast relative to the same issuance time than the NWSRFS forecasts. Therefore the comparison between the 2 sets of forecasts will use forecasts relative to the same events and the verification metrics will be computed from similar sample sizes.

The next teleconference will be on **Monday, May 5 at 1:30 pm EST**.