RFC Verification Workshop
Final Discussion
08/16/07

This report aims at summarizing the issues discussed by all the workshop participants at the end of the workshop on 08/16/2007.

1. Archiving / Hindcasting

All the RFCs should archive at least their operational forecasts (both deterministic and probabilistic). These forecasts integrate all the MODS, which are difficult to regenerate retrospectively. The impact of MODS on the forecast quality needs to be evaluated through rigorous verification studies using these operational forecasts (as well as other forecasts using different MODS). Given the efforts the forecasters put into the MODS process, the forecasters want to know whether and how MODS improve the forecast performance

Also the operational deterministic forecasts could be used as inputs for the HMOS work (part of the XEFS project) to generate probabilistic forecasts using statistical relationships between observations and operational forecasts.

Some RFCs (for example OHRFC or MARFC) are archiving the information (forecasts from different MODS scenarios or OFS files with or without MODS) to study the impact of MODS. Most RFCs archive their OFS files on a regular basis (but generally not on a daily basis).

So far, the archiving requirements for both deterministic and probabilistic forecast verification are not defined, although the Verification Plan developed in October 2006 included the need for a national archiving team to define these requirements. This work could start with a survey sent to the RFC Verification focal points to define what is actually available in their archive (type of forecasts – deterministic and probabilistic –, observations, location, time period, OFS files, forecast scenarios, etc.). This needs to be done in collaboration with the RFC Archive Database Update and Maintenance team. Additionally, most RFCs have reported issues with the current archive database, which have led to data lost.

Hindcasting (which is the process used to generate retrospectively forecasts for a fixed forecast scenario) is necessary to pinpoint the various sources of uncertainty in the different steps of the forecasting system. When evaluating the impact of the input forecasts on streamflow forecasts, the reference forecasts should be the streamflow generated from the perfect inputs. When evaluating the impact of the initial conditions, the forecaster should compare streamflow forecasts from different sets of initial conditions (VAR vs. no VAR for example). It is also important to account for the uncertainty from the rating curves. It would be useful if the forecasts of both streamflow and stage would be archived, as well as the rating curves (which may change with time).

The Deterministic Hindcaster should be developed and the Ensemble Hindcaster should be expanded to generate hindcasts from various forecasting scenarios. Meanwhile, the forecasters could generate different forecasts based on a range of scenarios to analyze the sources of skill and uncertainty. This work will be proposed for the WR and NWS Hydrology Forecast Verification teams.

2. Raw Model Forecasts / Other reference forecasts

Given the number of MODS used by the forecaster and the effort put into this operation (as the results from the verification workshop survey show), it seems extremely important to evaluate the impact of MODs (each MODS as well as different combinations of MODS) on the quality of the forecasts. It seems reasonable at this point to let each RFC define its own Raw Model forecasts since the definition varies a lot from one office to another. The WR and NWS Hydrology Forecast Verification teams could work on how raw model forecasts would be defined at each RFC and then use their raw model forecasts as a baseline to evaluate the forecast quality.

Besides, other reference forecasts (climatology, persistence, etc.) need to be used in the verification studies to evaluate the benefits of using forecasts produced by the forecasting system under evaluation vs. forecasts from other sources. Using a reference forecast (to compute skill scores for example) could be useful to evaluate whether the forecasts perform better because the events are more easily predictable (the reference forecasts would also perform better, so the forecast skill would not be high), or because of the “smarts” of the forecast system itself (the reference forecasts would not perform better, so the forecast skill would be higher).

3. Timing error vs. other errors

It is important to separate the timing error from the other types of errors (error in the peak value, shape of hydrograph, etc.). For some applications, even if the timing is not correct, the forecasts could be useful to the users. For example a flood warning could still be useful even if it is 12-hour off.

To analyze a set of forecasts, the first step is to visualize the forecasts and observed data to see the most important errors. Then different metrics could be used to analyze the errors. Most of the metrics will analyze the forecast variable for a given time step (for example 6-hour forecast), not the forecast time series.

To define a timing error, it is necessary to define the event from the observations and the forecasts, for which a timing error could be computed. The definition of events to know what the start time and end time of the event are (for both forecast and observed events) is a very complex process. STAT-Q is one of the prototypes developed at OHD to evaluate timing errors. However the definition of the events is not very robust since it is defined through threshold values or manually.

For the other types of errors, it is necessary to define a meaningful variable to be verified, such as minimum 7-day streamflow, or peak 30-day streamflow. Also it could be useful to compare the probability distribution functions of both forecasts and observations to detect some of the errors for a given verification time period.

4. Understanding of verification metrics

More material, exercises, and case studies are needed to understand all the verification metrics. The use of synthetic forecasts for which their errors are known (for example with a bias in the mean) could be useful to better understand how the results from the metrics look like for these types of errors. It is also important to explain the basics since it involves a lot of statistics. The WR and NWS Hydrology Forecast Verification teams will work on common verification exercises before the verification case studies to expand the forecasters' expertise on the verification metrics.

5. Verification software

The different verification applications (IVP, EVS, and logistical measures) seem very promising. All the RFCs are ready to work with these applications. Since no other tools offer the same capability, it is important to deliver these applications to the RFCs as soon as possible. Therefore the RFC verification focal points will more immediately apply what they have learned during the workshop and follow up on the hydrologic verification activities that have started during the workshop.

6. NWS Hydrology Forecast Verification team

This team should start working on exercises using IVP and EVS and share the verification results among the participants to improve the expertise of all the RFC verification focal points with the various metrics. Then each RFC could work on specific verification case studies and present their verification results to the team.

The agenda for the team should be consistent with the deployment of IVP and EVS to the different RFCs. The latest versions of IVP (to be released for

AWIPS ob8.2) and the EVS prototype (version 1.0) will probably be available to all the RFCs in December 2007. The agenda for the team will be modified according to this date.

7. Verification metrics for different users (forecasters, scientists, emergency managers, public, or NWS management)

Not the same verification metrics can be used for all the users since each user has a specific perspective on forecast quality and specific questions to be answered. The NWS Hydrology Forecast Verification team will work on defining verification strategies for specific users by analyzing verification case studies with identified end users.

For the NWS management, it is extremely difficult to summarize the forecast quality with only one number since there are multiple aspects to be evaluated in verification. There are a few metrics that are used internally (such as accuracy and lead time of detection). More verification case studies are needed to know what the most meaningful verification metrics are for specific users, including the NWS management (as the Verification System Requirements team recommended in October 2006). It will then be possible to develop useful verification metrics for the NWS management.

8. Training

There is a clear need for more training (sessions and materials). A similar verification workshop could be organized every year to know what progress is being made. Training could also involve go-to meeting to work on specific examples, as well as on-line modules (for example on specific software). Also the workshop participants are ready to use the workshop material (which is available on the workshop website) to train other forecasters at the RFC.

Matt Kelsch, the workshop participant from COMET, will develop a 1-hr module on hydrologic forecast verification using the workshop material, which will be available on line. COMET could also offer a longer (> 1 day) hydrologic verification session during the Advanced Hydrology course for the summer 2008. Additionally, verification training needs to be coordinated with the NOAA/NWS training team.