Towards Standard Verification Strategies For Operational Hydrologic Forecasting

Report of the NWS Hydrologic Forecast Verification Team

September 30, 2009

Team members

Julie Demargne (OHD/HSMB), Mary Mullusky (OCWWS/HSD),
Larry Lowe (ABRFC), James Coe (APRFC),
Kevin Werner, Brenda Alcorn and Lisa Holts (CBRFC),
Alan Takamoto (CNRFC), Kai Roth (LMRFC),
Bill Marosi and Andrew Philpott (MARFC),
Julie Meyer (MBRFC), Mike DeWeese and Holly Reckel (NCRFC),
Rob Shedd and Tom Econopouly (NERFC),
Joe Intermill and Stephen King (NWRFC), Tom Adams (OHRFC),
Christine McGehee (SERFC), and Greg Waller (WGRFC)

Verification technical advisors

James Brown (OHD/HSMB), Yuqiong Liu (OHD/HSMB) and Hank Herr (OHD/HSEB)

Team website

http://www.nws.noaa.gov/oh/rfcdev/projects/rfcHVT_chart.html

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Weather Service
Silver Spring, Maryland

Table of contents

Executive summary

All forecasts are unavoidably imperfect due to various kinds of errors and uncertainty involved in a forecasting process. Thus, forecast verification is necessary to evaluate the quality of the forecasts and to guide the development and improvement of a forecasting system. The purpose of this team report is to recommend standard strategies and products for hydrologic forecast verification to answer the following five overarching questions:

1) How good are the forecasts?

2) What are the strengths and weaknesses of the forecasts?

3) What are the sources of uncertainty and error in the forecasts?

4) How are new science and technology improving the forecasts?

5) What should be done to improve the forecasts?

The proposed verification standards consist of verification metrics and products for both single-valued and probabilistic forecasts, as well as verification analyses to be performed at all the NWS River Forecast Centers (RFC) for the five overarching questions. They reflect a consensus among the 13 RFCs, the NWS Office of Hydrologic Development (OHD), and the NWS Office of Climate, Water and Weather Services (OCWWS) on forecast verification for operational hydrology. These standards have been identified based on experiences with the verification software developed by OHD (Interactive Verification Program (IVP) and Ensemble Verification System (EVS)) and other existing verification capabilities (e.g., Western Region Water Supply Forecast website, National Precipitation Verification Unit, National Digital Forecast Database verification, and local applications at the RFCs), as well as on-going collaborations with scientists in the NWS, other hydrologic agencies, and the academia. These standards will all be implemented in the unified verification system that is currently under development within the Community Hydrologic Prediction System (CHPS) (Demargne *et al.*, 2009).

The report describes groups of forecast users and the type of verification information that will help them better utilize the forecasts. Given the variety of forecast applications and the different attributes of forecast quality, different levels of verification information, each containing several verification metrics and products, are needed to meet the needs of all users. In this report, four levels of information have been identified. Key verification metrics and products for both single-valued and probabilistic forecasts are presented for each level. Metrics are selected to facilitate the comparisons between the performance of single-valued forecast and that of probabilistic forecast, using similar metrics for both types of forecasts where possible (e.g., Mean Absolute Error or MAE for single-valued forecasts corresponding to Mean Continuous Rank Probability Score or CRPS for probabilistic forecasts). Below is a summary description of the four levels of information with key metrics (the table with all recommended metrics is given on page 22):

- Level 1: data display plots (time series plots and box and scatter plots) with the observed and forecast values used to compute verification statistics.

- Level 2: summary scores to be plotted for multiple forecast points on spatial maps, such as skill scores for MAE and Mean CRPS for single-valued and probabilistic forecasts respectively.
- Level 3: more detailed scores and sampling uncertainty information on verification metrics, such as the reliability and resolution components from the decomposition of Mean Square Error (MSE) and Mean CRPS, as well as Relative Operating Characteristic (ROC) Score as a discrimination measure.
- Level 4: sophisticated verification results to describe the forecast performance for specific event thresholds (e.g., above Flood Stage), using for example ROC curves as a discrimination measure, and False Alarm Ratio and Reliability diagram as reliability measures.

Verification analyses to be performed by modelers and forecasters are discussed for each of the overarching questions, together with recommendations to obtain reliable and meaningful verification results. These recommendations include:

- The impact of any newly developed forecast process (e.g., new calibration parameters, new preprocessing technique, new observed dataset) should be analyzed via systematic verification. When comparing two forecasting scenarios, the verification results need to be produced for the exact same events (e.g., same verification period, same time step). Therefore verification results should be reported separately for the daily forecast points and the flood only points.
- For forecast performance tracking purposes, it is necessary to define meaningful groups of basins to aggregate the verification summary scores without masking potential forecast improvement for individual forecast points. By working on verification studies at all the RFCs, the verification team plans to develop criteria (such as basin response time) for the definition of forecast groups that have similar hydrologic processes, which should therefore show similar improvement in verification scores.
- Spatial aggregation of verification results across different basins should be carefully performed, not to mask large variations of forecast performance among the basins. Verification statistics should be first analyzed for individual basins and plotted on spatial maps to define subsets of basins for which the verification results have similar characteristics.
- The use of normalized metrics (e.g., skill scores) and metrics defined for common probability thresholds (e.g., from the observed probability distribution), rather than absolute thresholds is necessary, especially when comparing verification results across different basins (e.g., they will have different flow magnitudes); it is also necessary when aggregating these verification results (under the condition that the basins show similar verification characteristics).
- Temporal aggregation is necessary to verify different forecast products at different time scales (e.g., 6-hourly instantaneous flow forecasts vs. weekly minimum flow forecasts). Although the 6-hourly temporal resolution is the primary scale for verification of operational forecasts (excluding water supply forecasts), it is recommended to define other

time scales to support specific users. The verification team should define a few forecast products for longer time scales to be verified at a national level.

- Forecasts should be first verified for each individual lead time since forecast performance varies greatly with lead time. Then, by analyzing the verification statistics as a function of lead time, one can assess whether to pool forecast data across multiple lead times if the verification metrics are similar for these lead times (i.e., pool the four 6-hourly forecasts for lead day 1 to produce verification statistics for lead day 1 at 6-hourly time step); the data pooling will increase the sample size but should be performed only if the verification statistics from the individual lead times are similar.
- Forecast performance should be analyzed under different conditions by stratifying the forecast-observed dataset. Data stratification should be based on both time conditioning (e.g., by month and by season) and atmospheric/hydrologic conditioning. For inter-comparison purposes, the verification team should agree on a few categories for data stratification to report verification results at a national level. It is recommended to define atmospheric/hydrologic conditioning based on low and high thresholds defined from the observed probability distribution, and on specific absolute thresholds (e.g., probability of precipitation, freezing level, and flooding level). It is important not to define too many categories for data stratification so that the sample size for each category contains enough data to give reliable verification statistics.
- Verification results should be reported along with the sample sizes since the sampling uncertainty could have a significant impact on the values of the verification statistics for small sample sizes (which is usually the case for extreme events). Work is underway to estimate and represent the sampling uncertainty in the verification metrics with confidence intervals. Once the verification software has the capability to estimate confidence intervals, it is recommended to report verification measures accompanied by the confidence intervals for a given confidence level (which will be defined at a national level to ensure homogeneity among the RFCs).
- The different sources of uncertainty and error need to be analyzed by verifying both the forcing input forecasts and the hydrologic outputs. For extreme events, both flow forecasts and stage forecasts should be verified since verification results of flow and stage could be significantly different due to the quality of the rating curves for such events. Sensitivity analysis of the different sources of uncertainty relies on using different forecasting scenarios. Two sensitivity analyses for single-valued stage forecasts are recommended to be performed at all RFCs: 1) impact of the QPF horizon on the hydrologic forecast performance, by using QPF forecasts of increasing horizons (e.g., from 6-hour to 5 days); 2) impact (on a day-to-day basis) of run-time modifications made on the fly on the hydrologic forecast performance, by using two forecasting scenarios with and without run-time modifications made on the fly (but including the a priori modifications, which are defined by the forecasters before producing any forecast). The verification team recommends a set of common baseline scenarios to be used at all RFCs for these two studies, although each RFC could define additional scenarios to meet specific local needs (e.g., stage forecasts produced from longer QPF horizon).

Proposed standard verification products are given for the four levels of information mentioned above, using IVP and EVS graphics, as well as additional plots presented by RFC forecasters and collaborators. These standard products are proposed to initiate discussions between RFC forecasters and external users on which verification products are the most meaningful. These verification products should be generated for the recommended verification analyses and for the four levels of information, although the first three levels (data display plots and verification scores on individual forecast points and on spatial maps) should be sufficient for most users.

Required enhancements of the current IVP and EVS software and verification science are identified; most of the proposed scientific enhancements were already included in the OHD verification activities for FY09. The main enhancements concern: 1) the analysis of timing error information of flow forecasts; 2) the estimation of confidence intervals (along with a graphical capability) to represent the sampling uncertainty of the verification metrics; 3) the consistency of verification information for weather and water forecasts; work is underway with NCEP to use similar verification metrics and report results for spatial areas that are consistent with the hydrological modeling performed by the RFCs (e.g., verification statistics for each RFC area). Additional efforts are needed for data archiving (which is crucial to archive all data and metadata required for verification), hindcasting (to retroactively generate forecasts from a given scenario with large enough sample size), as well as verification training for RFC forecasters and forecast users.

Finally future activities for the verification team are proposed to:

- Produce, evaluate and improve the verification standards with expanded verification case studies at all RFCs. In their verification case studies, the RFCs should determine which verification products would be the most meaningful for them and for their forecast users. They should also demonstrate how verification would help guide improvement of the forecasting system and the forecast process. The team will develop prototype functionalities to produce the verification standards with the existing software (IVP, EVS, WR water supply website, and the CHPS display capabilities). Such analysis will help define criteria to aggregate verification results across basins and track forecast performance on the identified groups of basins.
- Define what-if scenarios to specify which observed and forecast datasets, spatial and temporal scales, verification metrics and products should be used for a range of situations (e.g., drought forecasts, flood forecasts, record forecasts, and tidal forecasts) and for a range of applications.
- Perform detailed user analysis of the verification products in collaboration with the RFC Service Coordination Hydrologists and OCWWS and develop requirements for dissemination of verification information for RFC river forecasts by the NWS Performance Branch and by the RFCs (the verification products accompanying the forecast products).
- Continue to support the design and development of the CHPS Verification Service (CHPS-VS) by testing verification prototypes (e.g., EVS) and reporting requirements and necessary enhancements for a unified verification system that meets all user needs.

A second team charter is proposed in Appendix C to perform these future activities from October 2009 to September 2011.

The proposed verification standards are likely to evolve as new verification science and software are being developed, for example to account for the uncertainty in the observations, to verify extreme events and account for climate change, and to verify spatial and temporal joint distributions (not only forecasts at a single location for one specific lead time). Collaborative research work is under way with universities (e.g., University of Iowa, University of California, Irvine, Iowa State) and NCEP/EMC, as well as scientists involved in the Hydrologic Ensemble Prediction Experiment (HEPEX) verification test-bed (which involves Environment Canada and ECMWF). The role of the verification team (which includes all RFCs) seems essential to ensure that these collaborative efforts will lead to common verification products and practices for weather, climate and water forecasts, thus meeting the needs of all forecast users.

Introduction

The NWS Hydrologic Forecast Verification Team was chartered in July 2007 with the following team charter (available on the team website: http://www.nws.noaa.gov/oh/rfcdev/projects/rfcHVT_chart.html).

Vision: River forecast verification tools and information will be readily available to users including forecasters, service hydrologists, managers, and the general public. Verification information will be meaningful to each user group. RFC forecasters will generate and communicate river forecast verification results and identify shortcomings to be addressed through software, system, or information enhancements. Ultimately, forecast verification will be successful when its results help determine action items within each user group.

Statement of the Problem: Currently, information on NWS river forecast performance is limited in scope and generally not communicated to most user groups. In recent years, nationally supported verification software has been developed which has great potential to address user needs. However, this software remains largely untapped.

Mission: To communicate meaningful river forecast verification information to user groups including forecast users, forecasters, service hydrologists, and management using existing software (IVP and EVS). This mission includes three major components: (1) developing understanding of verification statistics and concepts, (2) developing expertise with IVP and EVS software, and (3) developing standardized verification strategies to effectively communicate results to identified end users while ensuring verification needs are met.

Success Criteria: The team will develop a final report by September, 2009 that proposes standardized verification strategies to effectively communicate verification results to identified end users. To accomplish this, each RFC focal point will write a brief report describing a verification case study that identifies a specific user group, presents river forecast verification results, and highlights unmet needs. The team leader will coordinate with the RFC verification focal points to ensure the verification case studies consider the broad spectrum of hydrologic products and users. Standard verification strategies will be identified through the case studies.

The team started to meet at the first RFC Verification Workshop on August 14-16 2007. Twelve teleconferences were held between November 29, 2007 and November 10, 2008 to discuss the archiving requirements and issues, work on two verification exercises (one with IVP and one with EVS), and review the RFC verification case studies (a short description of the 13 RFC verification case studies is given in Appendix A). The team met again at the second RFC Verification Workshop on November 18-20, 2008 to discuss progress on verification software and science, as well as RFC verification case studies and verification activities. All workshop material is available online at http://www.nws.noaa.gov/oh/rfcdev/projects/rfcHVT_workshop2_agenda_presentations.html.

The team interim report was made available on 01/22/2009 (http://www.nws.noaa.gov/oh/rfcdev/docs/NWS-Verification-Team_interim_report_Jan09.pdf).

The interim report includes a description of the data archiving requirements (which was delivered to the IWT Archive Team in May 2008), the 13 RFC verification case studies, and the recommendations and actions from the second verification workshop.

The team held five other teleconferences from February to September 2009 to develop standard verification strategies described in this team report. The proposed verification strategies have been identified based on experiences with the verification software (IVP and EVS) and other existing capabilities, such as the Western Region Water Supply Forecast website, the National Precipitation Verification Unit, the National Digital Forecast Database verification, and local applications at the RFCs. Also these strategies were based on on-going collaborations with scientists in the NWS, other hydrologic agencies, and academia (mainly Allen Bradley from the University of Iowa, Kristie Franz from Iowa State, Yuejian Zhu from NCEP, and Vincent Fortin from Environment Canada). The proposed verification standards will help develop a comprehensive verification service within the Community Hydrologic Prediction System (CHPS), which is called CHPS Verification Service (CHPS-VS) (Demargne et al., 2009).

This report is organized as follows. First it describes the different overarching questions in forecast verification and the different groups of users. Key verification metrics and products, as well as verification analyses are then discussed for the different overarching questions and groups of users (a glossary of verification metrics is provided in Appendix B). Examples of standard verification products are proposed and required enhancements of current verification applications and services are identified. Finally future activities for the verification team are proposed to produce, evaluate and improve the standard verification products with RFC verification case studies and to support the development of CHPS-VS.

Overarching questions in verification and groups of users

Overarching questions in forecast verification

All forecasts are unavoidably imperfect due to various kinds of errors and uncertainty involved in a forecasting process. Thus, forecast verification is necessary to improve the operational hydrologic forecast process and to communicate the forecast skill and uncertainty to all users for better decision making.

The main overarching questions in verification are:

1) How good are the forecasts?
    - Since forecast quality is multi-faceted, several verification metrics are needed to quantify the forecast quality.
    - Since forecasts are used by multiple users for various applications, verification information must be given with several levels of sophistication.
    - Since the definition of "goodness" varies with application, these should include measures of skill relative to baseline forecasts of the same variables.
2) What are the strengths and weaknesses of the forecasts?
    - The forecast quality varies in multiple ways, thus several conditions should be used to verify subsets of forecasts and understand in which cases the system performs well (or not so well).
    - To get meaningful verification results, several reference forecasts (e.g., persistence, climatology) should be used for comparison.
3) What are the sources of uncertainty and error in the forecasts?
    - This requires the analysis of both forcing inputs and hydrologic outputs and the use of multiple forecast scenarios to analyze the impact of potential error sources. A hindcasting capability needs to be available to produce for the different forecast scenarios large sample of forecasts and therefore robust verification statistics.
4) How are new science and technology improving the forecasts?
    - This requires comparing verification results from the current system with results from a new system to objectively quantify the change in the forecast performance due to using the new system.
    - It also requires performance measure tracking to evaluate the level of success of river forecasting over time.
    - Improvements related to observing networks (e.g., observation availability in space and time, measurement accuracy, observation uncertainty estimates) should also be analyzed since observations are ingested in hydrologic and hydraulic models and are also used to assess the quality of the forecasts (note that the error in the observations is currently not accounted for when computing the verification statistics).
5) What should be done to improve the forecasts?
    - This decision needs to be based on the verification analyses performed for the four overarching questions mentioned above.

Verification activity has value only if the information generated leads to a decision about the forecast system being verified. Therefore the various users of the forecast verification

information must be identified and the verification tools need to be flexible to meet the needs of all users. Also, it is critical to present the verification results to the users in a transparent and meaningful way to allow them to make informed decisions based on the verification results.

Examples of users verification needs

Here are two examples of users and their needs, which were provided by NWRFC.

1) WFOs/Emergency managers: in flooding conditions, the RFC products are used directly or indirectly by emergency managers through the WFOs. The emergency managers look for information on forecaster confidence and historical bias in the forecasts. Simple tables to answer the following questions would be helpful: (i) when a flood was forecasted, what proportion of times did the river actually reach flood stage vs. not reach flood stage? (ii) when no flood was forecasted, how often did it flood or not flood? (iii) if a point was forecasted to go above Major Flood 12 times in the last 10 years, how many times did it actually reach major flood threshold? Combinations using other lead-times and/or thresholds would provide useful information as well.
2) Shipping industry: some of the mainstem, hydrodynamic (tidally influenced) forecasts are used to make shipping schedules and cargo decisions, thus the users' ability to understand and trust our forecasts has a significant financial impact. The river levels change quite dramatically and very quickly throughout the two high tides and two low tides each day. These customers have expressed interest in knowing how "far off" our forecasts can be. The forecast errors can have different impact at peaks or troughs, as well as for different elevations (e.g., forecasts at low levels are extremely critical). Quick changes in the river stage make forecasts of peak/trough timing very critical as well. Tables showing forecast error in height for peaks and troughs throughout the gage profile would be useful. Incorporating timing errors would be useful as well.

Groups of users

In general verification supports:

- scientists/researchers and hydrologic program managers, by identifying needs to improve forecasting system and measuring the value of products from current and new science and technology;
- hydrologic forecasters, by defining acceptable methods to generate forecasts and products and satisfying user demands;
- emergency and water resources managers, and the general public, by quantifying forecast performance and uncertainty for better decision making.

Users of the RFC forecasts and the verification information include:

- RFC forecasters;
- NWS meteorologists;
- scientists and software engineers from NWS and the wider hydrologic community;
- partner agencies: WFOs, US Corps of Engineers, USGS, USACE, USDA, TVA, Coast Guard, reservoir operators, water managers, river gage operators;
- emergency managers, floodplain managers and organized ALERT groups;
- hydrology program managers;

- general public, including recreational users (who use, for example, snowmelt products for whitewater rafters), farmers, and the media;
- sophisticated users: power companies, transportation and shipping companies (e.g., barge industry), water supply managers, yield management companies.

Given the wide range of users, different levels of details are needed for the different applications. Also no single verification measure provides complete information about the forecast quality. Therefore different levels of sophistication are needed with several verification metrics and products to meet the needs of all users. The team agreed that the NWS should provide at least four levels of verification information, ranging from detailed statistics useful to forecasters, modelers and sophisticated users (e.g., with decision support tool that could directly ingest some verification results), to a few summary scores (e.g., with green-yellow-red code for each score) for the general public; one of the information levels consists of plots to display forecast and observed values since these data displays give the background information to understand the verification statistics. The NWS should also provide users with a lexicon that translates statistical results that are difficult to understand into understandable terms. A summary description of verification metrics with their strengths and weaknesses should be provided along with the results, to help users understand these statistics.

The different levels of information to provide the verification results for the various user groups could be summarized as follows:

- Level 1: display plots with the observed and forecast data used to produce verification results;
- Level 2: poor-fair-good color scaling of one or two summary score(s), likely to be sufficient for the general public;
- Level 3: additional summary scores and uncertainty bands on verification metrics (to describe sampling uncertainty) for emergency managers, program managers, recreation and agricultural users;
- Level 4: detailed statistics and case studies of extreme events for partner agencies and sophisticated users.

Due to the multiplicity of the applications based on the RFC forecasts and the need to conduct verification at various time and space scales, all the forecast and observation data should be made available to the users on the web, along with a verification application. The WR water supply website available at http://www.nwrfc.noaa.gov/westernwater/ gives an excellent example of such capability. This will enable sophisticated users to compute their own verification statistics based on their specific action or impact thresholds at the time and space scale of interest.

Types of verification information

Verification products provide two types of information:

- information on the quality of the delivered forecast services (called logistical verification), to evaluate the quality of forecast services in terms of the usability of the

forecasts (e.g., number of forecast locations, new type of forecasts, effort to issue forecast, forecast timeliness, etc.);

- information on the quality of forecasts, which includes verification of single-valued forecasts and probabilistic forecasts (ensemble and water supply forecasts) on different spatio-temporal scales (e.g., from hours and kilometers for flash flood guidance, to years and entire regions for water resource planning). This component needs to include diagnostic verification and real-time verification (Demargne *et al.* 2009). Diagnostic verification evaluates the quality of past forecasts given certain conditions (time period, variable value, event, methodology, etc.) to measure and improve model performance. It is done off-line with archived forecasts or hindcasts, which are sorted into different subsets according to specific conditions. Real-time verification evaluates the quality of live forecasts in real-time (before the observation occurs) using performance of past forecasts under the same or similar conditions as a guide to future performance, and should aid forecasters in making decisions when producing the forecasts. This may involve querying and displaying historic analogs to the real-time forecast, displaying summary of past verification statistics, checking up potential forecast anomalies, and if necessary, bias-correcting the live forecast.

This report proposes diagnostic verification standards, including summary diagnostic results and data display plots that could be meaningful to the forecasters for real-time verification also. Other requirements for real-time verification (e.g., the analog selection) are currently being identified in collaboration with the RFCs. Development of the real-time verification component is strongly linked to existing and planned capabilities in CHPS for forecast generation, analysis and product generation (e.g., the Graphics Generator prototype). Therefore requirements and software development for real-time verification will be carried out in collaboration with the OHD Hydrologic Science and Modeling Branch (HSMB) and Hydrologic Software and Engineering Branch (HSEB), Deltares, and the RFCs. Requirements for logistical verification will be developed in collaboration with OCWWS/HSD and the RFCs.

Key verification metrics and products for question 1: How good are the forecasts?

Attributes of forecast quality

Forecast quality includes several aspects or attributes. Forecast quality refers to the degree of correspondence between individual pairs of forecasts and observations of what actually occurred (or a good estimate of the true outcome). Attributes of the forecast quality include (Wilks, 2006):

- Bias in the mean (or first order bias, overall bias, unconditional bias), to analyze how the "best single-valued estimate" from the forecast agrees with the observed outcome on average. For single-valued forecasts, the best estimate is the forecast value itself; for probabilistic forecasts, it is generally the ensemble mean forecast, but could involve some other measure of central tendency, such as the median or mode.
- Correlation, to describe the linear relationship between forecasts and observations.
- Skill, to estimate if the verified forecast is more or less accurate than a given reference forecast. Skill requires the selection of one verification metric and one reference forecast, which is usually climatology, persistence, or random chance.
- Reliability, to describe the agreement between, for one or more subsamples of the verification data, the observations for the subsamples and the respective forecasts. It is relative to the conditional distribution of the observations given the forecasts and helps answer questions like: if a flood event was forecast, did it actually occur?
- Resolution, to describe the ability of the forecast to sort a set of observed events into different subsets with different frequency distributions. It is also relative to the conditional distribution of the observations given the forecasts.
- Discrimination, to describe whether the forecast system can discriminate between events and non-events; this is relative to the conditional distribution of the forecasts given the observations; it helps answer questions like: if the observations are in the flood level category, what did the forecasts predict?
- Sharpness for probabilistic forecasts (which is an attribute of the forecasts alone), which accounts for the need to make predictions with extreme probabilities (high or low), and not probabilities close to climatology.

For now, the observations are considered the "truth" even if they are imperfect (e.g., due to random errors, bias and representativeness errors in measurements, and analysis errors when the observational data are analyzed or altered to match the scale of the forecast). It is currently assumed that errors in the observations are much smaller than the errors in the forecasts.

Also the assumptions of stationarity in the observations and the forecast models (i.e., observations and forecasts are generated from the same processes) enable the users to pool forecast and observed values from different events and compute verification metrics to quantify the different quality aspects. Therefore whenever a forecast model changed, the forecasts generated with the new model should not be pooled with forecasts produced by the previous

model. To get large enough samples of forecasts, one needs to use a hindcasting capability to generate retroactively the forecasts based on the updated forecasting system.

Verification results need to be produced and inter-compared between single-valued forecasts and probabilistic forecasts for forcing inputs and hydrologic outputs. Therefore the verification metrics that could be computed for both types of forecasts would be preferable.

Key verification metrics for the different attributes of forecast quality

Given the verification literature and the experiences of the team with various verification applications, the following verification metrics are proposed as key metrics to analyze the different aspects in forecast quality and quantify the different forecast attributes. A glossary of verification metrics is given in Appendix B. Further details are available in Wilks (2006) and Jolliffe and Stephenson (2003).

1) For single-valued forecasts

- Forecast error:
  - Mean Absolute Error (MAE), which is generally preferable to MSE (and RMSE) because it is less sensitive to large errors from specific forecasts and it can directly be compared to CRPS for probabilistic forecasts;
  - MSE (and RMSE) can be offered as a secondary metric since MSE can be decomposed into the following three components:

    MSE = Reliability - Resolution + Variance (Obs),

    Here reliability is the bias conditioned on the forecast; variance (corresponding to 'uncertainty' for probabilistic forecasts) indicates the intrinsic difficulty in forecasting an event and is independent of the forecast system being evaluated.
- Bias:
  - Mean Error (ME) or relative measures such as the Relative Bias or Percent Bias; one should note that a bias defined as a ratio between forecast mean and observed mean could lead to misinterpretation for very small observed mean.
- Correlation:
  - Correlation Coefficient (CC)
- Skill:
  - MAE Skill Score computed with a given reference forecast (MAE-SS_{ref}) (see glossary of verification metrics for skill score definition);
  - MSE Skill Score (MSE-SS_{ref}) (and RMSE-SS_{ref}) can be offered as a secondary metric if users prefer MSE

- Reliability (conditioned on the forecast):
    - Reliability component in the MSE decomposition ($Reliability_{MSE}$) as a summary measure of reliability;
    - for a given outcome/event (e.g., flow above flood level), False Alarm Ratio computed from the 2x2 contingency table using the positive forecast category;
- Resolution (conditioned on the forecast):
    - Resolution component in the MSE decomposition ($Resolution_{MSE}$) as a summary measure of resolution
- Discrimination (conditioned on the observation):
    - for a given outcome/event (e.g., precipitation above 1 in), Probability of Detection POD and Probability of False Detection POFD, which are computed from the 2x2 contingency table using the observed categories;
    - Relative Operating Characteristic curve (plot of POD vs. POFD) to consider both metrics together for a given outcome/event; there is one curve for each set of forecast-observed pairs and for a given event; the forecast has some discrimination skill if its ROC curve is above the diagonal line (POD=POFD);
    - ROC Score as a summary score on ROC for a set of outcomes/events; ROC Score is derived from the area below the ROC curve (called ROC Area); the ROC Score is equal to zero if the ROC curve corresponds to the diagonal (POD=POFD), meaning the forecast has no discrimination skill.

The reference forecasts for skill score computation for single-valued forecast are:

- persistence, defined as the last observed value maintained for all lead times, except for temperature forecasts at sub-daily time steps; for example for 6-hourly temperature, the last four 6-hourly observed values will be used for each day to keep the diurnal cycle;
- climatology, which could be defined as the average observed value from the same date in the historical record; one could compute the average on a 30-day moving window centered on the given date to smooth these climatological values.

2) For probabilistic forecasts

- Forecast probability error:
    - Mean Continuous Rank Probability Score (Mean CRPS) to describe the average probability error (averaged across all forecast-observed pairs); the Mean CRPS corresponds to the MAE for single-valued forecasts; the CRPS can be decomposed into 3 components:

$$CRPS = Reliability - Resolution + Uncertainty(Obs);$$

    - Brier Score (BS) for a given threshold can be offered as a secondary metric to describe the average square probability error of the probability

forecast of a binary event; BS could be useful to users for whom a specific threshold is important (e.g., probability of precipitation or PoP, probability of flooding);

- Reliability (conditioned on the forecast):
  - Reliability component in the Mean CRPS decomposition ($Reliability_{CRPS}$) as a summary measure of reliability;
  - Cumulative Talagrand diagram to describe the distribution of the observations, on average, within the probability forecast; there is one curve for a given set of forecast-observed pairs;
  - Reliability diagram for a given outcome/event to plot the forecast probabilities vs. the relative observed frequencies, the forecast probabilities being divided into K bins; there is one curve for each set of forecast-observed pairs and for each event; the reliability diagram is also useful to assess the sharpness of the forecast for the specified event since it includes a histogram of forecast sample sizes vs. each of the K forecast probability bins;
- Resolution (conditioned on the forecast):
  - Resolution component in the Mean CRPS decomposition ($Resolution_{CRPS}$) as a summary measure of resolution;
- Discrimination (conditioned on the observation):
  - ROC plot (POD vs. POFD) for a given outcome/event and with N probability levels (binary classifiers) to turn the probabilistic forecast into a yes/no forecast; there is one curve for each set of forecast-observed pairs and for each given event; the ROC curves for single-valued forecasts and probabilistic forecasts for a given event can directly be inter-compared if plotted together;
  - ROC Score as a summary score for the ROC curve for a set of outcomes/events; ROC Scores can also be inter-compared for single-valued and probabilistic forecasts;
- Skill:
  - Mean CRPS Skill Score using a reference forecast ($CRPSS_{ref}$);
  - Brier Skill Score (BSS_{ref}) for a given threshold can be offered as a secondary metric to verify binary events;
- Deterministic measures:
  - To compare a specific aspect of the performance of probabilistic forecasts with that of single-valued forecasts, a few statistics (e.g., bias and correlation coefficient) can be computed from the "best estimate" derived from the ensemble forecast, which might be the ensemble mean. However these metrics reflect only the quality of the "best estimate" thus losing information about the distribution of the probabilistic forecasts.

Reference forecasts to compute skill score for probabilistic forecasts could be:

- climatology based on a long historical record, which is the most obvious reference probabilistic forecast; (a conditional climatological forecast could also be defined to provide a more skillful reference);
- "naïve" forecast that could be defined as a pseudo-persistent ensemble forecast, although such reference is not frequently used (note that persistence forecast cannot be used for probabilistic forecast verification since it is defined deterministically); one could use a fixed distribution (e.g., Normal for temperature) with the ensemble mean equal to the last observation and the standard deviation estimated from the "range of observations" for a given past period (e.g., last 3 months); the selection of distribution could be quite complex, especially for precipitation given its intermittency.

The metrics mentioned for single-valued and probabilistic forecasts reflect all the forecast errors when the forecast is paired with the observation according to its valid time. However the forecast error for streamflow should be characterized into timing error, peak error, and shape error to better understand the forecast quality (a similar characterization can also be done for precipitation forecast with timing, intensity and pattern errors). Such approach requires pairing forecasts with observations based on events. OHD is currently working on adapting techniques from spatial verification, curve registration, and cross wavelet analysis to develop these new error metrics. Also these verification metrics are relative to forecast points and do not account for any statistical dependencies in space, for which verification techniques will be investigated in the future.

Also, to complement the information on forecast quality, one could assess the economic value of forecasts for particular decision problems since it is expected that enhanced forecasts should result in better decision making and thus improve economic consequences. Even if the value of a set of forecasts will vary with the decision problem, the cost/loss ratio approach gives insight into the forecast value for simple real-world decision problems. One of the most common measures to measure forecast value is the Relative Value (or Economic Value) (which is used by NCEP for their ensemble forecast verification for example). It is a skill score of expected expense using a Cost/Loss ratio, with climatology as the reference (see http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html for references and Appendix B). Since the Relative Value depends on the Cost/Loss ratio, it is plotted as a curve for Cost/Loss ratio varying from 0 to 1 for a given event. Relative Value can be computed for both single-valued and probabilistic forecasts. As for any skill score, if Relative Value is greater than zero, the forecast has more potential value than climatology; otherwise the forecast is worse than climatology. One user could determine his/her own Cost/Loss ratio and analyzes how different sets of forecasts (single-valued and/or probabilistic) might help compared to climatology. Although the Relative Value is not available in IVP and EVS, OHD plans to include it in the unified verification system in the future.

Normalized verification metrics

Normalized verification metrics are necessary to inter-compare or aggregate verification results at different forecast points and across different RFCs. They take into account the differences between verification results due to the basin characteristics and the type of outcome/event that

the forecast predicts. For example, the same mean error value in flow forecasts for a large basin will not impact the forecast quality in the same way as it does in a small basin; instead, the forecast for the small basin tends to be more 'biased', given its smaller range of flow values. Also, by comparing to the same reference forecast, a normalized metric helps determine if the verification results are good because of the real 'smarts' of the forecast system or because the event is easier to predict.

Skill scores are one example of normalized metrics. They are based on the selection of one verification metric (e.g., MAE and Mean CRPS) and one reference (persistence, climatology, or chance). For single-valued forecast, persistence is usually defined as the reference (e.g., in IVP); for probabilistic forecasts, climatology is usually used as a reference. The choice of the reference depends on the application. Comparison with persistence indicates how well the model predicts changes; persistence is likely to be more difficult to beat for shorter lead time. Comparison with climatology indicates how well the forecast performs in unusual situation and climatology usually performs well at longer lead time. Note that climatology should be defined from a record longer than the sample climatology from the verification dataset because it is more stable. Verification results for climatology forecasts should indicate how these forecasts were defined. Additional references could be defined in the future as it was suggested in the previous section.

Other relative measures can be defined. For bias, Relative Bias or Percent Bias (see glossary for formulas) are usually chosen. Regarding the Nash Sutcliffe Efficiency (which has been extensively used to evaluate the calibration of hydrologic models), several articles question the utility of it because it assumes mean observed value as benchmark, over-emphasizes large values, and has a minimum of negative infinity. Therefore the metric is not currently included in the recommended verification metrics.

Also specific thresholds or outcomes for verification computation (e.g., ROC) should be defined to facilitate inter-comparisons of verification statistics. The thresholds can be defined as percentiles in the observed distribution (e.g., 10^{th} percentile, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, 90^{th} percentile). One should note that these percentiles are tied to a probability distribution (e.g., empirical distribution) and will have a different meaning (or real value) depending on that distribution. To inter-compare results from various basins, the same probability distribution should be used to define these percentiles (for example, the empirical distribution). The specific impact thresholds are still needed to verify the forecasts for individual basins (e.g., Action Stage, Flood Stage). However when inter-comparing the forecast performance for multiple basins, the verification results based on the basin impact thresholds are more difficult to interpret since these extreme events could be significantly different across basins (e.g., with very different sample sizes or observed frequencies).

Aggregating verification results across different basins can be meaningful when using normalized metrics and when the basins have similar hydrologic regime (e.g., similar response time) but should be performed carefully. For example, the MARFC case study with EVS (see Appendix A) demonstrated how the verification statistics could highly depend on the basin response time. Therefore the verification metrics should be computed first for individual forecast points and then plotted on a spatial map to analyze the homogeneity (or non homogeneity) of the

verification results across multiple forecast points. Aggregating the verification statistics across different basins is meaningful if these basins show similar verification statistics.
Besides, the verification statistics need to be first computed and presented for each individual lead time since the forecast quality varies greatly with forecast lead time. It is necessary to analyze the verification statistics as a function of lead time. If the verification metrics are similar for different lead times, one could pool the forecast data across these lead times. For example, one could pool the four 6-hour forecasts for lead day 1 to produce verification statistics for lead day 1. The data pooling will increase the sample size but should be performed only if the verification statistics from the pooled dataset will not mask significant variations of the verification statistics at the individual lead times. Such data pooling across lead time is feasible in the operational IVP ob8.3. One of the recommended IVP enhancements listed in this report is to include in the GUI a recommendation to first analyze the verification statistics at individual lead time in order to check whether data pooling across lead times would be meaningful.

One should note that data pooling across different lead times is different from temporal aggregation, which defines a new variable with a different time step from that of the original forecasts. For example, daily maximum flow may be defined from 6-hourly instantaneous flow for both forecasts and observations to perform verification of this new forecast variable. Temporal aggregation depends on the forecast application and should include various statistics, such as minimum, maximum, average, and total. A flexible functionality to perform temporal aggregation before verification needs to be included in the verification system, as it is in the current EVS prototype (but not IVP).

Key metrics and products for the four levels of verification information

Forecast and verification information have multiple users. Therefore different levels of sophistication are required for the verification metrics and plots. Different levels of spatial aggregation should be provided to give verification information for individual forecast points and for groups of forecast points. Also different time periods are required to provide verification information from the last “days”, from specific past events of interest, and from the last “years”. Some of the proposed products for diagnostic verification could also be used for real-time verification, but this report focuses on diagnostic verification products only.

First, it is necessary to provide data plots, to show the forecast and observed data used to compute verification metrics. Useful information can be gathered from the data display plots. This is especially true if the user is analyzing forecasts relative to a single past event, generally for a few days (see the CNRFC and NWRFC verification case studies for the analysis of forcing inputs and flow forecast quality for a single storm event). For such analysis, the visual comparison of the forecast values with the observed values could be sufficient to analyze the forecast performance. Note that, since the metrics for each specific lead time are computed based on a very few forecast-observed pairs for a single past event, the forecaster needs to generalize the findings by looking at similar past events (as in the real-time verification approach): by selecting historical analogs from the past, the forecaster could detect systematic errors in the subset of observed-forecast pairs, which could be relative to specific conditions. Such study is part of the verification analyses to be done for the overarching questions 2 and 3, which are described later in this report.

The data displays are relative to both single-valued and probabilistic forecasts. They should include plots of forecast values vs. observed values and plots of forecast error values vs. observed values. For probabilistic forecasts, each forecast time series needs to be plotted with box and whiskers to represent the distribution of the probabilistic forecast values (the user defines which probability levels should be used). Data display plots need to be provided in the following situations:

- when providing the most recent operational forecast: a time series plot should display all the operational forecasts issued for the last 5 to 10 days and the observed time series; such plot gives some insight on the forecast agreement with the observations and the forecast consistency (i.e., how close the forecast values are from two consecutive issuance dates) from the recent past; it also gives a sense of the forecast uncertainty (similarly to the approach with time-lagged ensembles) and how difficult it was in the past few days to predict the future conditions;

- when providing background information for historical events (these events could be selected as historical analogs by the forecaster in real-time verification): similarly to the plots for the operational forecast, the time series plots would give for each historical event the forecast and observed values for all the forecasts issued during a 5 to 10 days time period, or longer (depending on how long the event of interest lasts); this information will help the user better utilize a real-time forecast if the forecast is similar to the historical events;

- when providing verification results for a given set of observed-forecast pairs: scatter plots would represent for each individual lead time the observed-forecast pairs used to compute the verification metrics. For single-valued forecasts, each forecast-observed pair is represented by one point, whereas for probabilistic forecasts, the probabilistic forecast values are represented by box and whiskers. Scatter plots could also be used to display forecast errors (i.e., forecast value – observed value) (for ensemble forecasts, the error is computed for each ensemble forecast value). This type of plot could be very useful to detect bias that is conditioned on the observed value (see examples later in this report).

In addition to the proposed time series plots, a flexible functionality has been developed by the CNRFC and made available in their Historical Graphical River Forecast Interface at http://www.cnrfc.noaa.gov/histRVFinterface_loop.php. It provides an animation of the forecast and observed time series for both rain and melt and stage/flow for a location, a forecast date and a forecast cycle selected by the user. Thus, the user has a quick and simple means of comparing the shape and timing to peak between the observed and forecast time series as well as viewing the observed and forecast rain and melt at the same time. Such functionality should be made available at all RFCs to enable users to derive time series plots for any forecast and observed data stored in the historical database.

The team agreed to propose the following verification metrics and plots for the four different information levels.

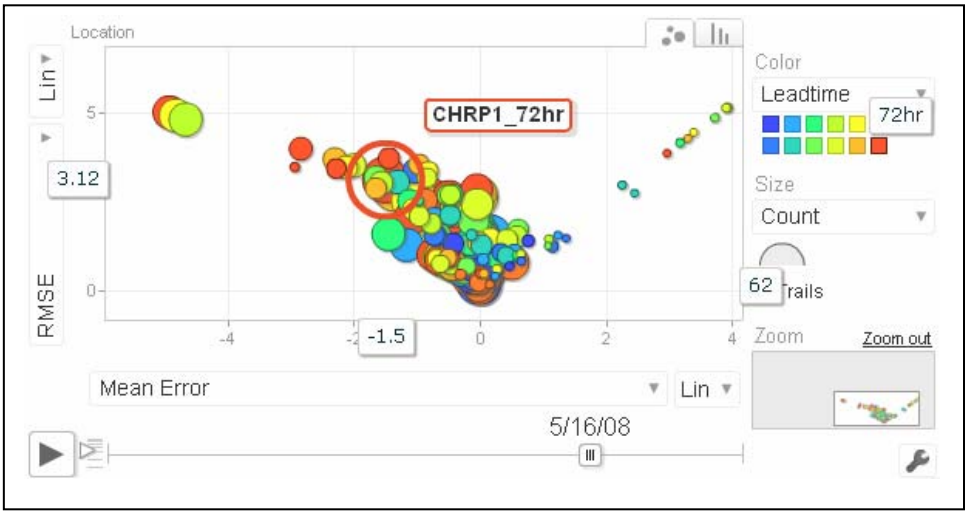
Information level	*Attributes*	*Single-valued forecasts*	*Probabilistic forecasts*
1) Data information	Forecast and observed values	Scatter plots for each lead time	Scatter plots with box and whiskers for each lead time
		Time series plots for set of forecasts	Time series plots with box and whiskers for set of forecasts
2) Summary information	Error	MAE	Mean CRPS
	Bias	Relative Bias	Relative Bias in ensemble means
	Skill	MAE-SS_{ref}	$CRPSS_{ref}$
3) More detailed information	Error	MSE	BS for set of events
	Skill	MSE-SS_{ref}	BSS_{ref} for set of events
	Reliability	$Reliability_{MSE}$	$Reliability_{CRPS}$
	Resolution	$Resolution_{MSE}$	$Resolution_{CRPS}$
	Discrimination	ROC Score for set of events	ROC Score for set of events
	Correlation	Correlation coefficient	Correlation coefficient for ensemble means
	Sample size	Number of forecast-observed pairs	Number of forecast-observed pairs
4) Sophisticated information	Reliability	FAR for set of events	Cumulative Talagrand Diagram, Reliability Diagram for set of events
	Discrimination	ROC curves for set of events	ROC curves for set of events
	Forecast value	Relative Value	Relative Value

Table of the four levels of information with recommended verification metrics

As mentioned before, one of the goals is to easily compare results between single-valued and probabilistic forecasts. ROC, ROC Score, and Relative Value are computed for both single-valued and probabilistic forecasts, and MAE and Mean CRPS are mathematically comparable.

There are different approaches for providing graphics with the summary information that reflects the different attributes of forecast quality. Since forecast quality has multiple facets, a few key metrics and scores need to be included, as recommended for the summary information level. Using a single metric or score as summary information for all users does not seem reasonable. To provide this summary information, the first approach is to plot the different metrics in one single plot to describe the various forecast attributes in one unified graphic. The second approach is to combine different metrics into one score, although working with a single score could be difficult: one needs to define what relative weights to use for the individual metrics and how to interpret the score since it integrates the information from different metrics. This approach will be investigated by OHD in the future.

Regarding verification graphics, the proposed approach for now will be to provide individual graphics for the recommended key metrics and scores, as well as combine several metrics into one single plot (examples of recommended graphics are given later in this report). In the verification literature, one example on how to present several verification measures into one graphic is the Taylor diagram (Taylor, 2001), which OHD plans to evaluate in the future. Such information could be presented with a 2-D plot, giving on the y axis, values for different metrics, as a function of lead times or time periods (seasons or months). Here is another example with the bubble plot developed by OHRFC (using GoogleMotionChart) and available at http://www.erh.noaa.gov/ohrfc/bubbles.php. On top of the two metrics that can be plotted on the x and y axis, other metrics can be represented by the color and the size of the points. In the example below, the Mean Error is plotted on the x axis, the RMSE on the y axis, the size of the circles represent the sample size, and the colors correspond to the different lead times. Such plot could also be animated to show how metrics vary in time.



Example of bubble plot provided on the OHRFC website

In conclusion, key verification metrics have been identified for both single-valued and probabilistic forecasts along with recommended verification plots for the four different information levels to describe the different attributes of the forecast quality. These verification products should be produced in the verification analyses described hereafter to better understand the strengths and weaknesses of the forecasts, the sources of uncertainty, and how new science and technology improve the forecast performance.

Key verification metrics and products for question 2: What are the strengths and weaknesses of the forecasts?

Data stratification

Forecast quality varies in time based on the different atmospheric and hydrologic conditions. One of the key questions in verification is to analyze when the forecasts perform well or not. Therefore the definition of different subsets of forecasts and observations to be verified will help assess how the forecast quality varies in particular conditions.

There are two main types of data stratification or conditioning (which could also be combined):

- time conditioning: subsets of forecasts are verified for different time periods (e.g., seasons, months);
- atmospheric and/or hydrologic conditioning: subsets of forecasts are defined for given conditions of precipitation, temperature, flow and/or stage observed or forecast values (e.g., high flow category if observed flow >= X). For example, by using a condition on either the forecast value or the observed value, the user can analyze the reliability and discrimination aspects of the forecasts. When using a threshold value for the forecasts, the verification results describe the reliability of the forecasts, whereas a threshold value for the observations will describe the forecast discrimination. The same verification metrics computed from two subsets conditioned on forecast value and observed value respectively will give complementary information (e.g., verification results for single-valued stage forecasts in Welles, 2005). One could also define more complex subsets of observed-forecast pairs, such as observed-forecast flow pairs when input forecast precipitation >= P.

Other data stratification could be useful as well, such as verifying QPF forecasts produced by each individual forecaster as done in the NERFC verification case study (see Appendix A); this could help forecasters identify systematic error in their forecasts and improve the forecast quality.

Regarding time conditioning, the recommendation is to produce verification statistics and plots for each month and for each 3-month season. In order to inter-compare across RFCs, the monthly and seasonal statistics should be produced for the same time windows (except for Alaska due the specificity of its hydrologic regime). Monthly statistics would be produced from the whole month. Seasons should be defined (at all RFCs except Alaska) as: December-January-February, March-April-May, June-July-August, and September-October-November. Additional time conditioning relevant to individual RFCs is also recommended (e.g., seasons defined according to the navigation seasons; Alaska may use only two seasons). The monthly and seasonal statistics can be aggregated for various years and/or for various basins if the basins have the same hydrologic regime (see the discussion on spatial aggregation on page 32).

For the atmospheric and/or hydrologic conditioning, multiple categories could be defined using either one or several variables. In order to inter-compare across RFCs, all the RFCs should agree on a few categories for routine forecast verification at a national level:

- categories defined from specific percentiles (of the sample distribution): 10^{th} percentile, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, and 90^{th} percentile are suggested;
- categories defined from specific impact thresholds: Action Stage and Flood Stage for flow and stage forecasts; probability of precipitation (PoP) for precipitation; freezing level for temperature (meaningful mostly for the Western RFCs).

Note that the WMO in 2008 recommended for the verification of single-valued and probabilistic precipitation forecasts the following rainfall intensity thresholds: 1, 2, 5, 10, 20, 50 mm/d (WMO, 2008). It is therefore recommended to include these thresholds in precipitation forecast verification. The use of other threshold values relevant to individual RFCs is also recommended.

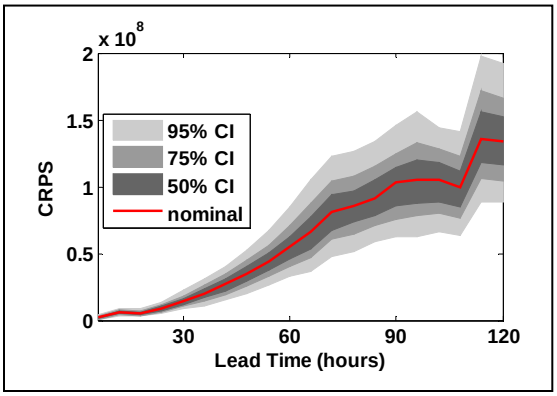
Sampling uncertainty

Since verification is based on finite samples of forecast-observation pairs, verification statistics are subject to sampling errors. The two key questions are:

- what is the uncertainty associated with the value of a verification measure?
- is forecast ***A*** significantly different from forecast ***B*** given the sampling uncertainty?

Sampling uncertainty in verification metrics needs to be estimated because sampling uncertainty increases with lead time and increases with decreasing sample size. The distribution of sampling uncertainty may change with sample size and is not always normal. When stratifying the verification dataset based on different conditions, sample uncertainty becomes an even bigger concern. Therefore the verification plots need to include the information about the number of forecast-observation pairs for each specific condition. Although data pooling may reduce sampling uncertainty, it is not recommended that data be pooled together from different lead times to increase the sample size unless the user has checked that the forecast quality is similar for the individual lead times. One way to account for sampling uncertainty in a specific verification metric is to compute confidence intervals, which are random intervals with a specified level of confidence (e.g., 95%, as recommended in (WMO, 2008)) of including a sample value of the metric (note that confidence intervals contain more information about the sampling uncertainty than a simple significance test). Verification plots with confidence intervals will likely to be meaningful to sophisticated users, who can integrate the information of the verification statistic uncertainty.

Work on sampling uncertainty is underway to compute confidence intervals for the different verification metrics using bootstrapping techniques. Examples of verification plots with confidence intervals were shown at the 2^{nd} RFC verification workshop in November 2008, including this example.



Example of Mean CRPS values vs. lead times with confidence intervals (CI) for three levels of confidence

In conclusion, different data stratifications are recommended to evaluate the forecast performance under different conditions. Data stratification should include time conditioning (by months and by seasons) and atmospheric/hydrologic conditioning, using both thresholds defined from percentiles (of the sample distribution) and specific impact thresholds (e.g., Flood Stage). The verification results need to be reported along with the sample size, and, in the future, with confidence intervals for a given confidence level.

Key verification analyses for question 3: What are the sources of uncertainty and error in the forecasts?

Analysis of the sources of uncertainty and error

Uncertainty in hydrologic predictions can come from different sources: the forcing inputs, the initial conditions, the model parameters, and the model structure. These uncertainties are typically referred to as either meteorological uncertainty (in the case of the forcing inputs) or hydrologic uncertainty (for all other sources). Hydrologic forecasts need to capture these two types of uncertainty and yet, they need to be as close to the observed outcome as possible (i.e., with small error) and with a better performance than the performance of a "naïve" alternative forecast (i.e., with skill). The relative importance of the meteorological uncertainty and the hydrologic uncertainty could vary greatly with basin characteristics, lead time, and hydrologic conditions, as well as with the spatial and temporal scales of the forecasts. These two uncertainty sources lead to two main sources of potential error, namely errors in the hydrologic model and errors in the atmospheric forcing. These two sources of error could interact with each other: the atmospheric error may attenuate or exaggerate the impact of the hydrologic error in the hydrologic forecasts.

For forecasters and modelers, it is necessary to analyze how the different sources of uncertainty impact the quality of hydrologic forecasts and which parts of the forecasting system represent the main sources of skill and error in these forecasts. To identify the sources of skill and error, all the forcing input and hydrologic output should be verified. Any verification study should include verification results of precipitation, and temperature (or other forcing input) if used in the hydrologic model, as well as verification results of hydrologic forecasts. For extreme events, it is recommended to verify both flow and stage forecasts. Rating curves are one of the models in the forecasting system to convert flow into stage, thus verification results for flow forecasts and for stage forecasts could vary significantly for extreme events.

For the forecaster, it could be very useful to perform a post-event analysis: for a single event, analyze the different sources of skill and error to better understand the forecast performance for that specific event (see in Appendix A the NWRFC verification case study for such analysis). However, findings relative to one single event need to be generalized to other similar events to understand how to improve the forecast quality in the future. This analysis can be done by using different forecasting scenarios (e.g., different QPF forecasts as done in the CBRFC case study, see in Appendix A) and inter-comparing the verification results. For example, Welles and Sorooshian (2009) analyzed the impact of the QPF forecast, the state updating technique, and the calibration parameters on the performance of NWSRFS single-valued stage forecasts. Such analysis is easier to set up with a hindcasting capability to retroactively generate two sets of forecasts from two different forecasting scenarios with large sample sizes.

When comparing forecasts from two different forecast scenarios, it is very important to use the same dates for the forecast issuance time to verify forecasts for the same events. This could be difficult if the two sets of forecasts are produced in real-time (without any hindcasting) and if the two forecast systems issue forecasts at different times of the day. For example, the WGRFC case study (see in Appendix A) presented a comparison between forecasts produced with variational

data assimilation (VAR) at every hour and non-VAR forecasts produced at every 6 hours. Such a comparison should include only the forecasts at 6-hourly time steps that are available in both datasets. If the verification samples are different, verification results need to be reported separately for each dataset.

Further work is necessary to better diagnose the sources of skill and error in both single-valued and probabilistic forecasts. In this report, two initial studies are recommended for all the RFCs: a QPF analysis (which could be extended to other forcing inputs), and a run-time modifications analysis.

Impact of QPF forecast horizon

Regarding the QPF analysis, the goal is to analyze what the optimized QPF horizon would be for hydrologic forecasts, i.e., how many lead times of QPF should be used to drive the hydrologic models. The choice of QPF horizon varies greatly across the 13 RFCs and could also vary within a single RFC area, with seasons, and with atmospheric conditions as the QPF quality varies greatly. Such decision should be based on rigorous criteria using verification results that would help evaluate in which situations a longer QPF horizon could be used to improve flow/stage forecasts. Therefore the verification team has been working on setting up at all the RFCs an analysis to inter-compare the quality of stage single-valued forecasts that are produced with different QPF horizons. This study would include a set of common baseline scenarios to be used at all RFCs, although it is recommended to define additional scenarios at each office to meet specific local needs. The requirements listed below are consistent with the QPF horizon study that is currently being conducted at MBRFC and NCRFC as part of the Central Regional verification effort.

This study requires producing and archiving in parallel various forecast runs, each one using a different QPF horizon, and evaluating in IVP the different sets of forecasts. A set of QPF horizons would be used for the different forecast runs. It is recommended to use the following QPF horizons at all RFCs: 0, 6 hours, 12 hours, 18 hours, 24 hours, 30 hours, 36 hours, 48 hours, and 72 hours. The QPF horizons of 96 hours and 120 hours are also highly desirable (even if these QPF products are currently delivered to the RFCs at a later time than the QPF products for shorter horizons). For the RFCs using longer QPF horizons for their operational forecasts, it is recommended to select a few of the short-term horizons and expand to longer horizons. Each RFC will select the best available QPF source for each forecast horizon. All the other components of the forecasting system (e.g., model parameters) should be identical to the ones used for operational forecasts, with the exception of the model states. These forecast runs should not include any run-time mods that are made on the fly by the forecasters, and run-time mods that could impact or interact with the QPF values used to drive the hydrological models. The run-time mods included in these runs need to be stored in metadata for all these forecast sets. The forecast to be produced and verified should be the 6-hourly stage on a 7-day window; the window should be longer for slow response basins. For each resulting stage forecast set, the verification metrics should be computed for each individual 6-hour lead time and for the whole verification period, as well as sub-periods relative to specific atmospheric or hydrologic conditions (e.g., hurricane season from May to November).

Additionally it could be useful to verify the stage forecasts with both the observed values and the simulated values, the simulated values being generated from the observed inputs using the same initial conditions and the same model (see in Appendix A the CBRFC verification case study with the analysis of QPF impact on forecast performance). Even though the impacts of the meteorological uncertainty and the hydrologic uncertainty on the hydrologic forecasts interact with each other, such analysis gives some insight into the relative impact of the two sources of uncertainty: impact of meteorological and hydrologic uncertainties when verifying with the observed values, and impact of the meteorological uncertainty (given the existing hydrologic uncertainty) when verifying with simulated values.

This QPF horizon study could be extended to other forcing inputs, or to evaluate forcing inputs from different sources and/or different forecasters (as done in the NERFC case study, see in Appendix A). Given that the impacts of the meteorological uncertainty and of the hydrologic uncertainty on the hydrologic forecasts interact with each other, this type of forcing input uncertainty analysis needs to include the verification of both forcing inputs and hydrologic outputs. However such analysis will not provide insights on the impact of meteorological uncertainty on hydrologic forecasts that are free of hydrologic uncertainty. In other words, with this kind of uncertainty analysis, one cannot completely separate out the influences of the two uncertainty sources.

This QPF horizon study is first proposed for single-valued forecasts. Similar studies will be needed in the future for probabilistic forecasting (and are already under way at OHD with experimental ensembles produced for the Hydrologic Ensemble Forecast Service) to determine which forcing inputs produce the optimized probabilistic hydrologic forecasts.

Impact of run-time modifications

The second analysis concerns the impact of the modifications done in real-time by the forecasters, or run-time mods. Such analysis would be very useful to the forecasters, who want to evaluate how much value they add to the forecasts in various forecast situations (see in Appendix A the NWRFC verification case study on the analysis of individual mods made during a specific flood event). It would also help program managers evaluate what should be done to improve the forecasts if one could identify key run-time mods (e.g., developing data assimilation technique to mimic key run-time mods and potentially improve the forecast quality). The analysis of the impact of model states on forecast quality may include many different aspects: for example, past modifications that impact the initialization of the hydrologic models, or modifications done on the fly to modify some model parameters or inputs. Even if the modifications performed at each RFC are very specific to each office, the team has agreed to set up a baseline model at all RFCs to assess the impact of run-time mods on the forecast on a daily basis (this analysis is similar to the OHRFC verification case study, see Appendix A). The goal is to assess the impact of the run-time mods made on the fly vs. the impact of all run-time mods on the quality of the hydrologic forecasts. This is based on the differentiation between two types of run-time mods: the a priori mods that can be defined a priori by the forecaster before producing any forecast vs. the mods that the forecaster makes on the fly by analyzing the current forecast. For example the a priori mods would include the mods relative to regulated points. These a priori mods would be included in all the forecast scenarios.

The study requires the definition of the reference model states that initialize the hydrologic models for all the forecast scenarios. The reference model states recommended by the team are defined by using the carryover saved 5 days prior to the current date. These model states integrate past manual modifications made by the forecasters but not the most recent ones, since the purpose is to assess the value added by the forecasters with run-time mods on a day-to-day basis. Stage forecasts are generated using:

- best available observed inputs (but no forecast input), with and without on-the-fly mods;
- best available observed and forecast inputs, with and without on-the-fly mods.

Such inter-comparison allows analyzing the impact of the run-time mods and their potential interaction with the forcing inputs. It is recommended to store metadata with the list of all the run-time mods being included in these runs. The forecast to be produced and verified should be the 6-hourly stage on a 7-day window; the window should be longer for slow response basins.

Further discussion among RFC forecasters is underway to agree on the a priori mods and to establish this standard baseline model nationally at all RFCs. It is also recommended to define other baseline models at each RFC to further analyze the impact of the different modifications done by the forecasters, since some of these modifications are very specific to each RFC.

In conclusion, it is recommended to analyze the different sources of uncertainty and error in the forecasts by inter-comparing multiple forecast scenarios. The verification results need to be compared using forecasts issued for the same events. They also should include both the forcing input forecasts and the hydrologic outputs. Two initial studies are proposed for all the RFCs to evaluate the optimized QPF horizon and the impact of the run-time mods made on the fly for the single-valued stage forecasts. The verification team is currently discussing how to implement at all RFCs the forecasting scenarios for these two verification analyses and is sharing current RFC experiences and scripts (e.g., at MBRFC, NCRFC, and OHRFC) to produce and store outputs from multiple forecasting scenarios. The analyses should be performed on a few forecast points that are representative of the RFC area, and for a minimum of one year (although multiple years would be required to get verification results for extreme events).

Given the workload to set up the various forecast scenarios and archive all the output data, and given the current CHPS implementation schedule, the team agreed that each RFC may set up their sensitivity analyses at different times, inside or outside CHPS. While some products could be generated with IVP and EVS outside CHPS, a subset of the verification products would be produced using the CHPS capabilities (Graphics Generator, FEWS Time Series Display, and FEWS Spatial Display). Therefore the implementation schedule would be different for the CAT RFCs and the CAT-II RFCs. First the CAT RFCs would start developing the verification standards using the CHPS display capabilities, and share their progress with the CAT-II RFCs. The CAT-II RFCs would develop the standard verification products using CHPS when their CHPS implementation is being finalized.

Recommendations for questions 4 and 5: How are new science and technology improving the forecasts? What should be done to improve the forecasts?

Verification in operational hydrology

Forecast verification should play a key role in operational hydrology. The impact of any (newly developed) forecasting process on forecast quality needs to be demonstrated via rigorous and objective verification studies prior to its operational implementation. Verification results should form the basis for accepting (or rejecting) proposed improvements to the forecasting system and the forecast process and for prioritizing system development and enhancements. The APRFC verification case study (see Appendix A) is an example of evaluating the impact of two calibration strategies on flow forecast performance.

Driving the operational hydrology research and development activities with verification results will be easier in the future when verification standards are agreed upon, when a unified verification system for both single-valued and probabilistic forecasts is available in CHPS, and when scientists and forecasters are trained on verification.

To evaluate what should be done to improve the forecasts, scientists, modelers and program managers need to rely on verification studies done for the overarching questions 2, 3 and 4. Analyzing the impact of the different forecasting steps or components on the forecast performance will help decide which components need the most improvement in a cost-effective strategy. This includes testing new observed datasets, which could potentially improve the forecast quality due to higher resolution in space and/or time, improved measurement accuracy and estimation of the observation uncertainty.

Forecast performance tracking

Forecast performance tracking is one important focus for program managers. Program managers need to monitor the forecast quality over time and show improvement in the forecasts. To do so, verification results are aggregated across large areas and inter-compared across RFCs. One example is the verification program supported by the NWS Performance Branch on single-valued stage forecast verification, which includes monthly statistics for RMSE, ME, and MAE. Here are a few suggestions on the verification information to be produced for such purpose.

For high level summary information, the most appropriate metrics are the skill scores for MAE and Mean CRPS. The skill scores should be computed for both climatology and persistence as reference. To show results for different regimes, the skill scores could be computed for a few subsets of forecast-observed pairs. The thresholds to define these subsets should be defined as percentiles in the observed empirical distribution (e.g., 10^{th} percentile, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, and 90^{th} percentile as suggested for question 2). For tracking the change in forecast reliability, FAR and the reliability component of CRPS could be used for single-valued forecasts and probabilistic forecasts, respectively. Regarding forecast discrimination, the ROC Score could be used for both single-valued and probabilistic forecasts. The reliability and discrimination measures should be based on a set of events defined as

percentiles (e.g., 10th percentile, 25th percentile, 50th percentile, 75th percentile, and 90th percentile).

One could expect to show in the long term some improvement by plotting the verification results for each year (see the MBRFC verification case study in Appendix A), even if the improvements were made at some specific points in time. The verification results should also be analyzed for each season since there could be different enhancements made to improve forecasts for each season (e.g., snow melt improvement). However verification statistics have sampling uncertainty; some of the variations in the metrics will be due to sampling uncertainty, and these variations could mask an improvement in the forecasting system. Therefore sample sizes (and in the future, confidence intervals for one specified level of confidence) should be provided.

The process of aggregating verification results across different basins while showing forecast improvement could be difficult since verification results vary greatly with basin characteristics (as noted in the OHRFC and MARFC verification case studies) and with atmospheric and hydrologic conditions (see in Appendix A the LMRFC verification case study describing how hurricanes could dramatically change the hydrologic behavior of specific forecast points). One of the criteria is the basin response time; basins are classified as typically fast, medium and slow responders (although the basin response time could vary with the atmospheric conditions for example). It may be necessary to define more subsets of forecast points to show improvement in specific forecast components. For example, regulated points, for which the uncertainty from reservoir operations could potentially mask the improvements made in the other forecasting components, may be treated in a separated group. Forecast points impacted by snow-melt could also be verified in a separated group. Also verification results for the daily forecast points and the flood-only points (see the SERFC verification case study in Appendix A) should not be aggregated since they do not correspond to the same observations and will have very different sample sizes.

The definition of meaningful groups of basins to track forecast performance should actually be derived from verification results from the overarching questions 2, 3, and 4 to group basins with similar characteristics. Based on their verification studies, each RFC should propose criteria to define subsets of forecast points that have similar verification characteristics. The verification team should then discuss the different criteria proposed by the 13 RFCs and agree on a few subsets of forecast points to track forecast performance at the national level.

In conclusion, forecast verification should be routinely carried out to analyze any new forecast process and to guide targeted improvements of the forecasting system and the forecast process. For performance tracking purpose, the verification results need to be reported for a few key metrics relative to the different quality attributes and for groups of basins. The definition of these basin groups, for which the spatial aggregation of the verification results is performed, should be defined by the RFCs based on their verification studies.

Proposed standard verification products

By reviewing the verification analysis for the five overarching questions, key verification metrics and plots relative to four different levels of information have been identified. In this section, verification product examples are given using various applications (IVP, EVS, WR water supply website) and various sources (e.g., products developed by the RFCs, products presented at the 2nd RFC verification workshop, products developed by NWS and Environment Canada). Along with the proposed graphical products, text products (not included here) should also be produced to describe the observed data (e.g., source, location, and original time step), the forecast data (e.g., name of model/forecast scenario, model initialization time, time step, forecast horizon), the verification products (e.g., verification method, verification period, lead time, accumulation period, spatial scale, thresholds, reference forecast, sample sizes), as well as provide the numerical reports of verification metrics.

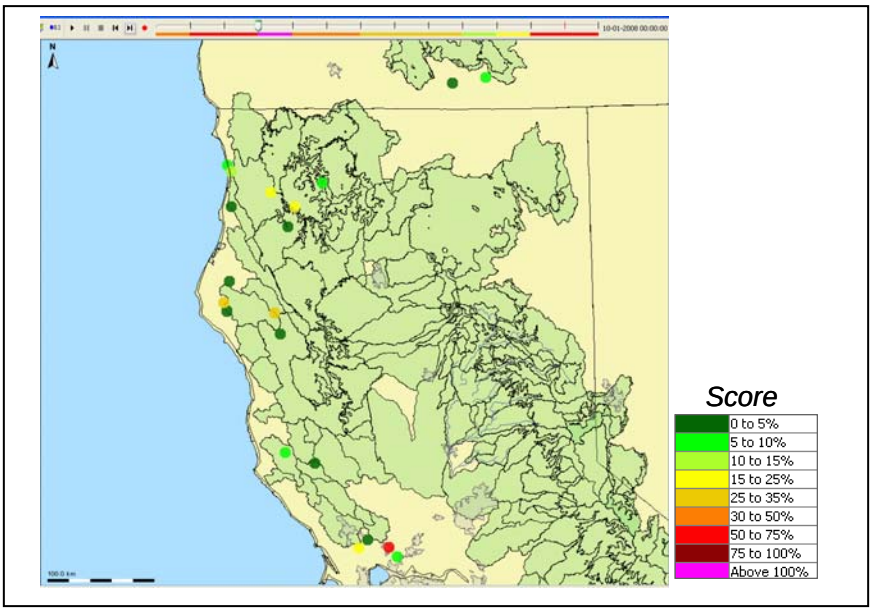
These examples are proposed as initial standards and are meant to initiate the discussion with the RFC forecasters and external users, about which verification products are the most meaningful. These proposed standards will have to be further evaluated by all RFCs with verification case studies. Further user analysis will also be needed to identify standard verification products for specific user groups. This analysis will be done with the NWS Verification Team, the CAT and CAT-II Teams, the Graphics Generator Requirements Team, and the RFC Service Coordination Hydrologists.

All information levels - Summary verification maps

As the first entry point for verification information, a summary verification map will display the value of a summary verification metric (e.g., MAE-SS_{ref}) for a given lead time and given time period for all forecast points. Map symbols will be used to indicate whether the verification metric is below or above a user-defined threshold. Symbols could be color coded (e.g., green-yellow-red code), use size, or some other iconic representation to indicate desirable or undesirable outcomes. These types of maps could be generated for any verification metric.

Maps are an excellent tool to analyze how verification varies with location. With animated maps, the user can also analyze how the verification statistics vary for different lead times, time periods (e.g., season or month), and thresholds. By clicking on one forecast point, the user could access more detailed information: data display plots, summary verification statistics, and more detailed or sophisticated statistics.

Such verification maps have been demonstrated by the HSMB at the CAT workshop of June 2009, using the FEWS Spatial Display. One example is given below; such map can be animated to show for example how the monthly values of the verification score vary in time. These verification maps as a function of lead times, time periods, and thresholds, should be made available in CHPS for all verification statistics.



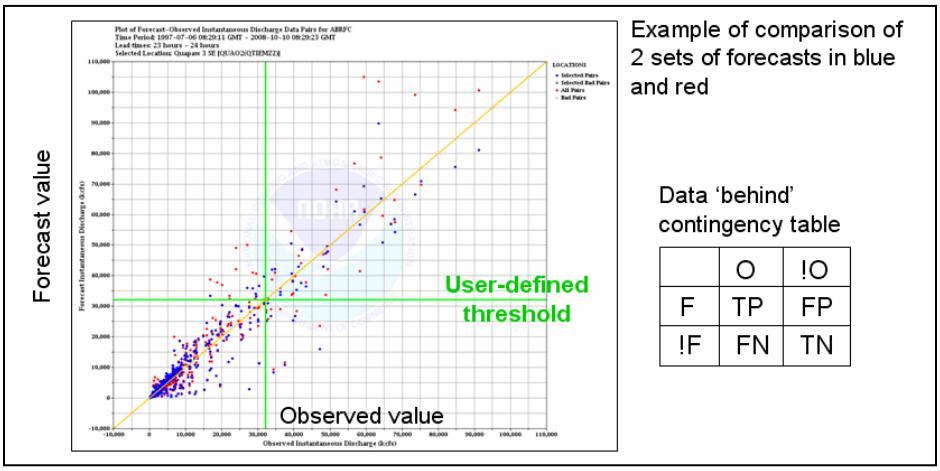
Example of summary verification map where a verification score is plotted for different categories using the FEWS Spatial Display

The NDFD verification website (http://www.weather.gov/ndfd/verification) includes several maps for which the user selects the variable (e.g., maximum temperature, PoP), the metric (e.g., MAE, Bias, Brier Score), the forecast cycle (0 Z or 12 Z), the forecast period (month), the forecast lead time, and the animation option (with lead time). For each map, the aggregated summary statistics are given for the whole area (i.e., all forecast points) and for four main sub-regions. Such functionality could be provided for forcing inputs and hydrologic outputs in CHPS-VS in the future.

Level 1 - Data display plots

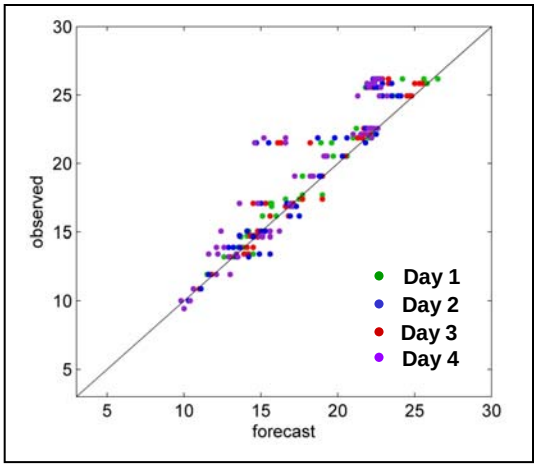
Here are examples of the two types of data display plots, scatter plots and time series plots, for both single-valued and probabilistic forecasts.

Scatter plots for a given lead time: IVP example for single-valued forecasts. By adding a specific threshold, the user displays all the forecast-observed pairs that define the 2x2 contingency table. In this example, the user can compare the quality of two sets of forecasts.



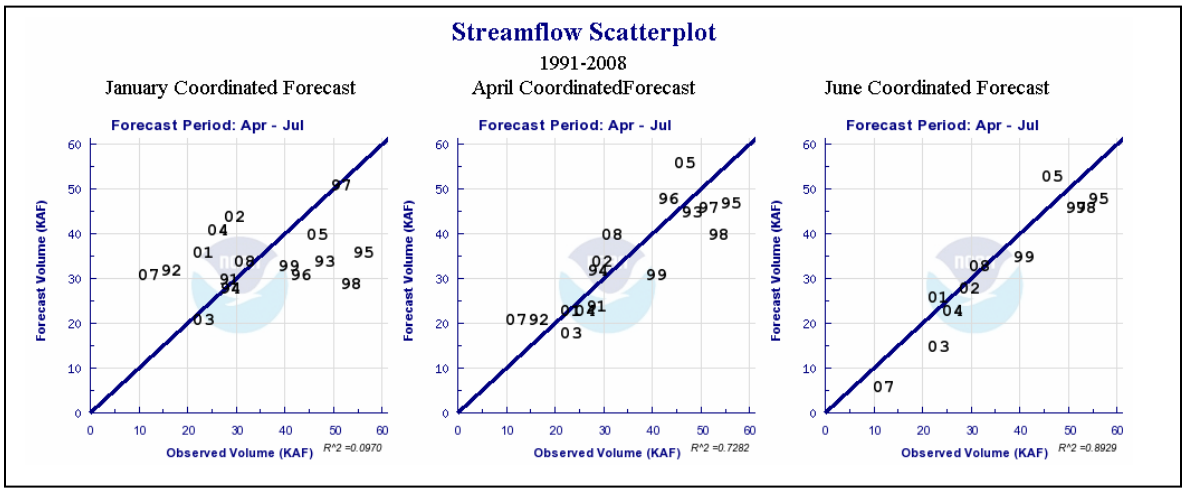
Example of scatter plot with a user-defined threshold

Scatter plots for a set of lead times: example from Kristie Franz (from 2nd RFC verification workshop). By using different colors for different lead times, the user can analyze how the forecast quality varies with lead time.



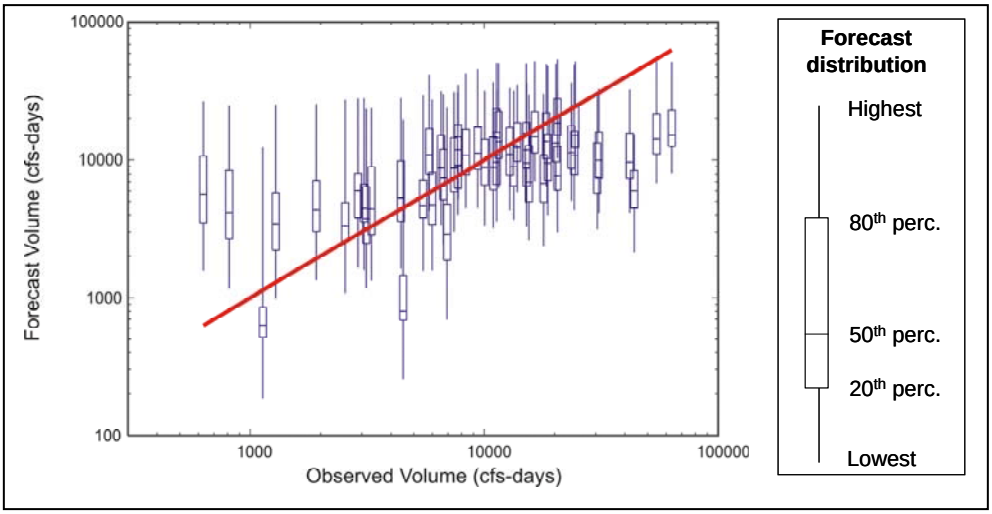
Example of scatter plot for a set of lead times

Scatter plots for a given set of historical forecasts: example from the WR water supply website. The user can analyze how the historical water supply forecast values vary with the forecast issuance time (in this case, January, April, and June).



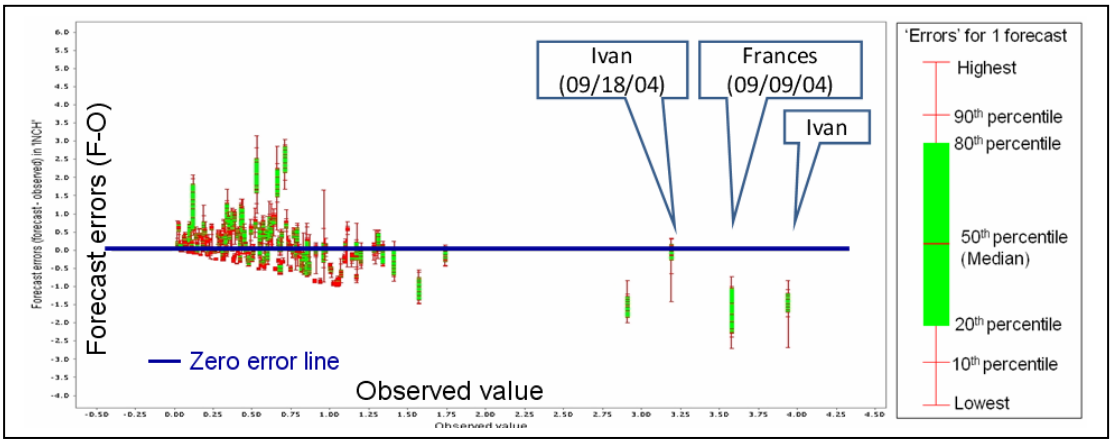
Example of scatter plot for a set of lead times

Box and whisker scatter plots for probabilistic forecasts for a given lead time: example of Allen Bradley (from 2nd RFC verification workshop). This is the equivalent plot for probabilistic forecasts to the first scatter plot for single-valued forecasts. Each probabilistic forecast is described with box and whiskers for given percentiles (in this case, 20th, 50th and 80th percentiles using the probability of non exceedance). The user can directly analyze how well the forecast values correspond to the observed distribution. For reliable forecasts, each of the box-and-whiskers forecast will cross the diagonal line.



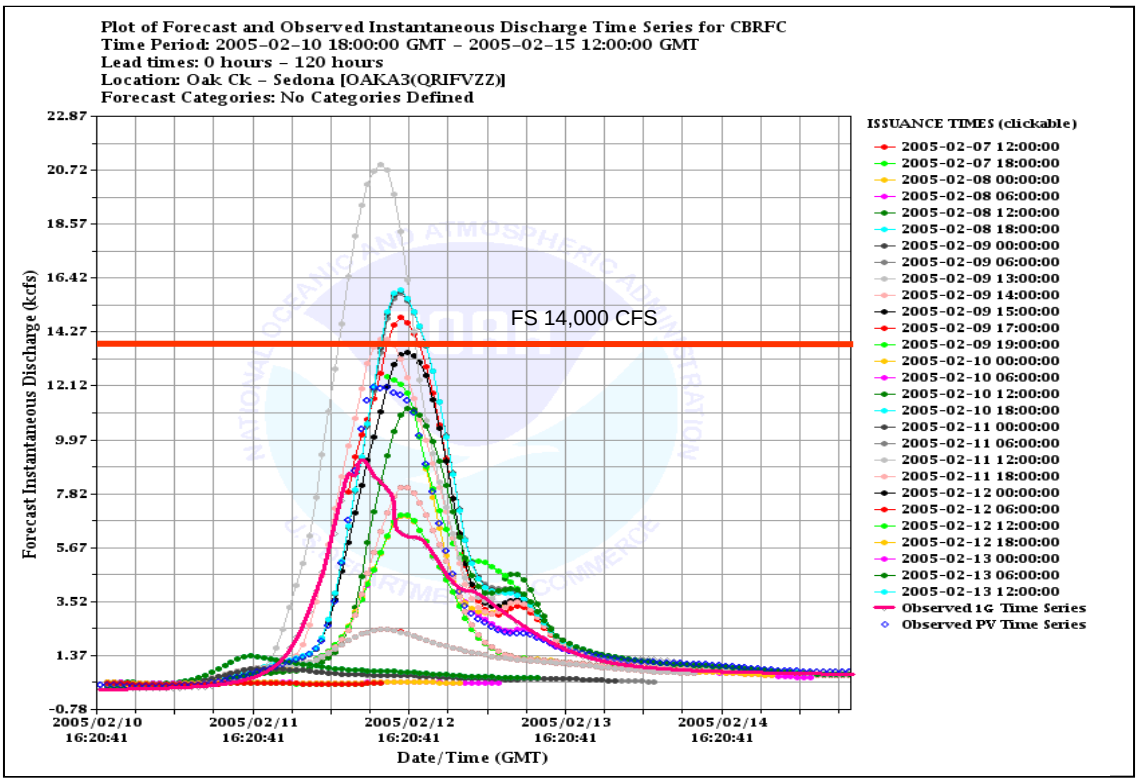
Example of box and whisker scatter plot for a given lead time

Box and whisker scatter plots of the forecast errors: EVS example. By plotting the forecast error as a function of the observed value, the user can detect a conditional bias. The labels with a short description of the event (in this example, hurricanes) would be also very useful for the user.



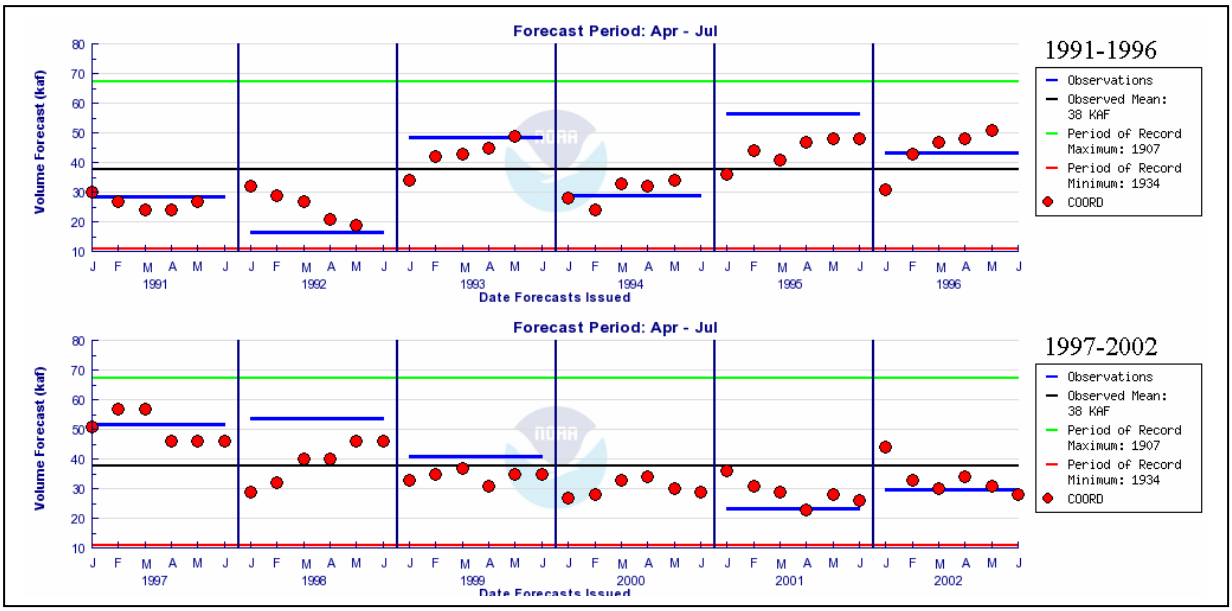
Example of box and whisker scatter plot for the forecast error, with the identification of historical events

Time series plots for a given set of forecast issuance times: IVP example from CBRFC verification case study. To analyze the forecast performance for one specific event, the time series plot is very powerful since the user can visually estimate the timing and magnitude errors in the peak values, as well as the rising and falling limbs.



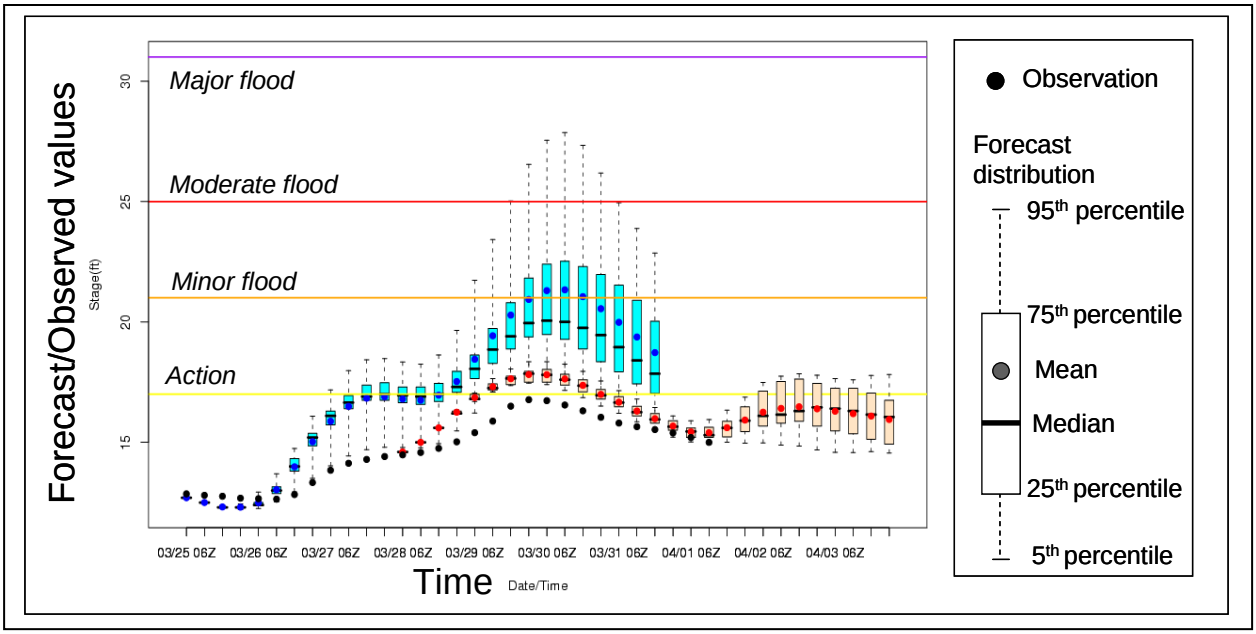
Example of time series plot for given forecast issuance times

Water supply historical plots for a given range of years: from the WR water supply website. The user can analyze how the forecast performance varies with the issuance time for each of the historical years.



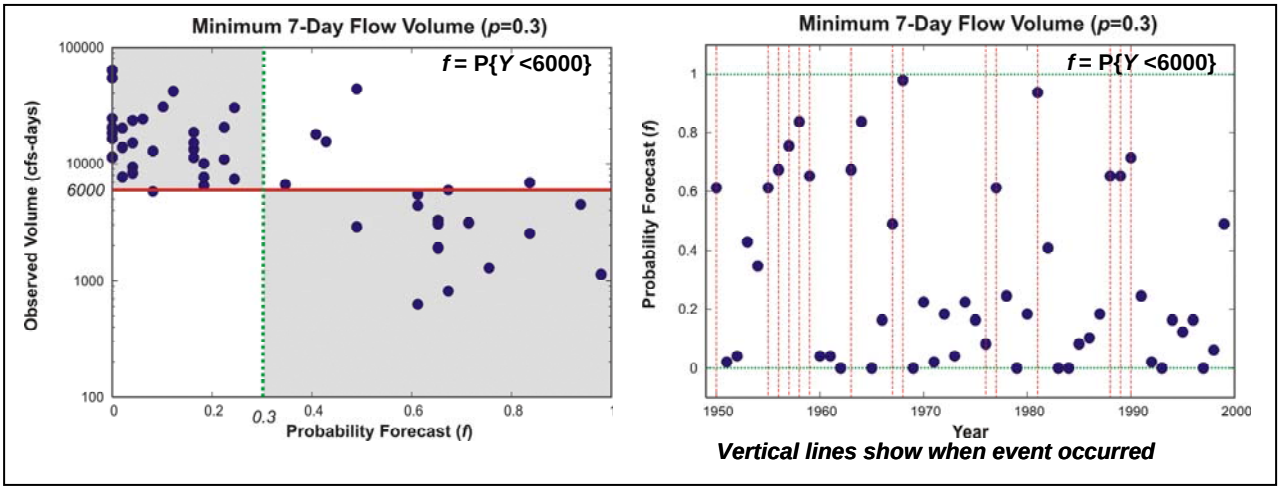
Example of water supply historical plot for the coordinated forecast (called COORD)

Box and whisker time series plots for probabilistic forecasts: example provided by OHRFC. In this example, the probabilistic forecast is represented with box and whiskers for five percentiles (5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles with the probability of non exceedance) with the ensemble mean overlaid. The time series of the probabilistic forecasts for two different issuance times (plotted in two different colors) can directly be compared to the observations. For reliable forecasts, the observations should be part of the forecast distribution. Note that such plot is readable only if the number of overlaid probabilistic forecasts is small. The plot readability may be improved if the representation of the forecast distribution is simplified (e.g., only median and whiskers for two percentiles).



Example of box and whisker time series plot for probabilistic forecasts

When the user is interested in a specific event, probabilistic forecasts could be transformed into event forecasts and the user can plot the observations vs. the probabilistic forecasts for the given event. Here are two examples of plots for a given specific event (flow volume < 6,000 cfs-day) from Allen Bradley's presentations at the 2nd RFC verification workshop. For perfect forecasts, when the event occurred, i.e., the observed value < 6,000 cfs-day (which corresponds to points below the horizontal line on the left plot and points with the vertical lines on the right plot), the forecast probability should be 1 (close to 1 for good forecasts); otherwise it should be 0 (close to 0 for good forecasts). Additionally, the user could transform single-valued forecasts into event forecasts (with forecast probabilities equal to 0 or 1). These two plots could then enable the user to directly compare the performance of probabilistic forecasts and single-valued forecasts.

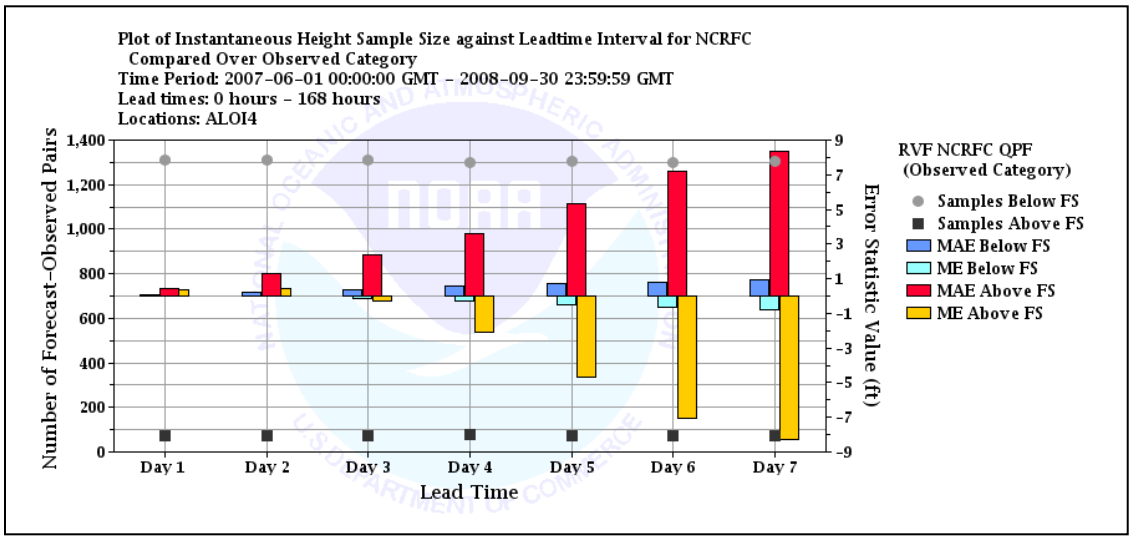


Examples of plots of probabilistic forecasts and observations for a specific event (flow volume < 6000 cfs-day)

Levels 2 and 3 - Verification statistical plots

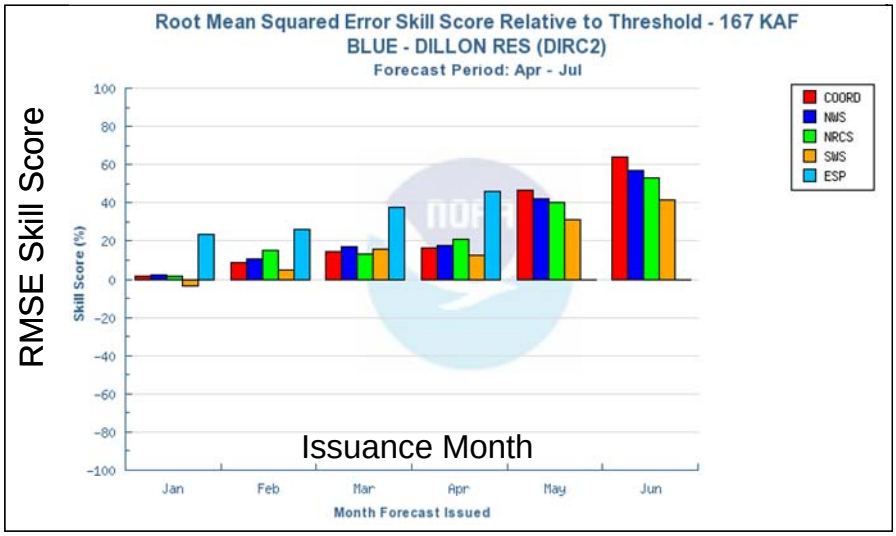
Plots of verification statistics should be presented in many different ways: for multiple seasons/months, for multiple lead times, for multiple models, for multiple thresholds. Sample size should also be provided along with the verification results. Here are a few examples.

Plot of metric values and sample size values vs. lead times for one or several subset(s) of single-valued forecasts: IVP example from NCRFC verification case study. In this case, the user can compare the metric values for two subsets of forecast-observed pairs (in this example, when the observations are above Flood Stage, and when the observations are below Flood Stage).



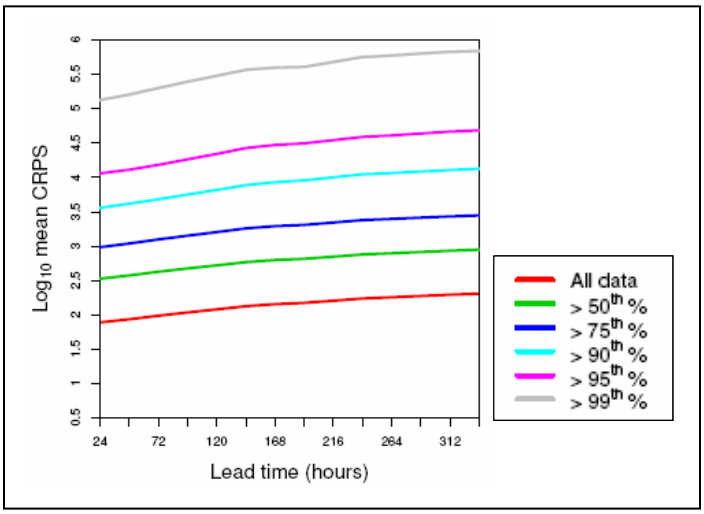
Example of MAE and sample size plot for two subsets of forecast-observed pairs

Plot of metric values vs. forecast issuance dates for one or several set(s) of forecasts: example for the WR water supply website. In this case, the user can compare the metric values for five sets of forecasts given the issuance forecast date.



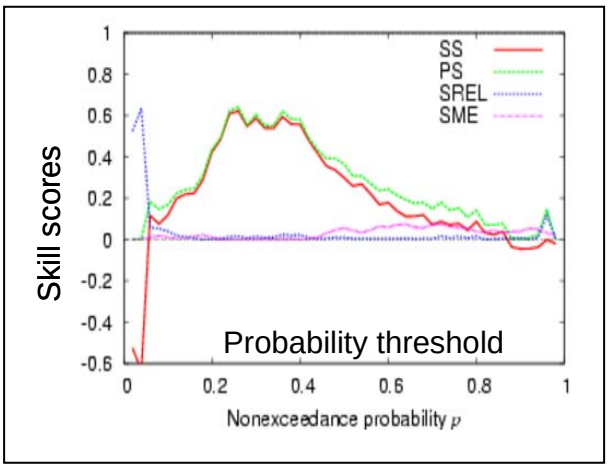
Example of RMSE skill score plot for several water supply forecasts as a function of the issuance month

Plot of metric values vs. lead time for different subsets of forecast-observed pairs for probabilistic forecasts: EVS example with the Mean CRPS values as a function of lead time. In this case, the five subsets of forecast-observed pairs are based on thresholds defined as percentiles from the observed distribution.



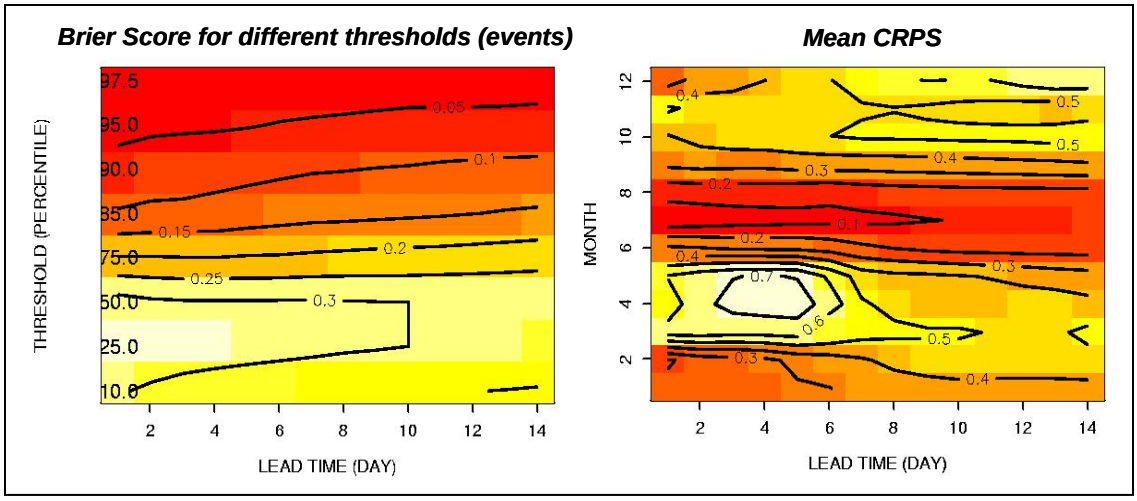
Example plot of Mean CRPS vs. lead time for several subsets of forecast-observed pairs

Plot of metric values vs. probability threshold: example from Allen Bradley (from the 2nd RFC verification workshop). In this example, the skill scores are plotted as a continuous function of the probability threshold. Such representation enables the user to define his/her own probability threshold of interest.



Example plot of skill scores vs. probability threshold

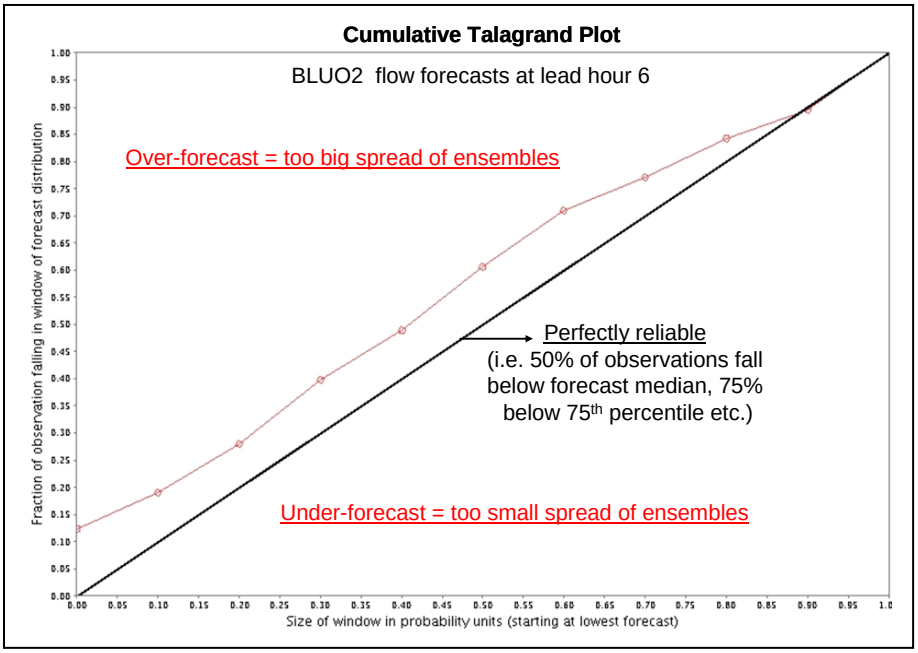
2-D plots of metric values relative to two different variables: EVS examples (from the previous version of the prototype). With such a plot, the user can analyze how the metric values varies with lead times and with the threshold values (on the left) or the forecast time periods (on the right).



Example plots of Brier Score and Mean CRPS as a function of lead time and of threshold values (left) and forecast time period (right)

Level 4 - Advanced verification statistical plots

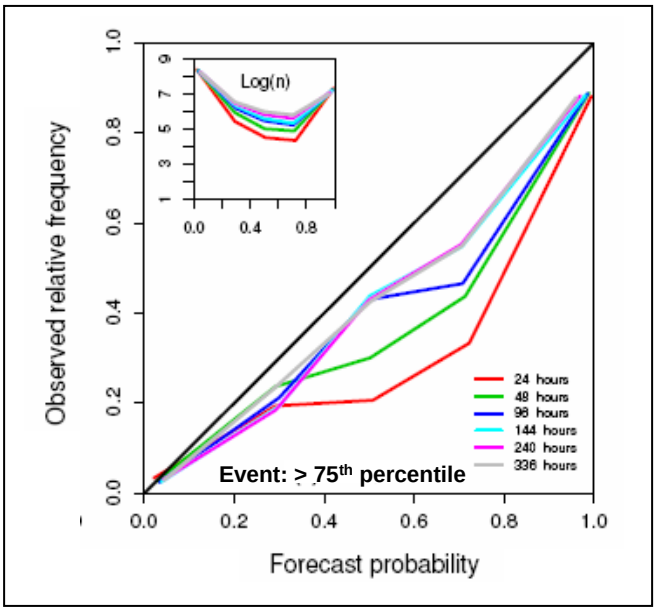
Plot of Cumulative Talagrand Diagram (similar to Rank Histogram) for a specific lead time: EVS example from ABRFC verification case study with plot labels to help users interpret the results. For perfectly reliable forecasts, the line should overlay the diagonal line. In this case, the ensembles are overspread or underconfident. (Note that such measure is more difficult to interpret for bounded variables such as precipitation).



Example of Cumulative Talagrand plot with labels

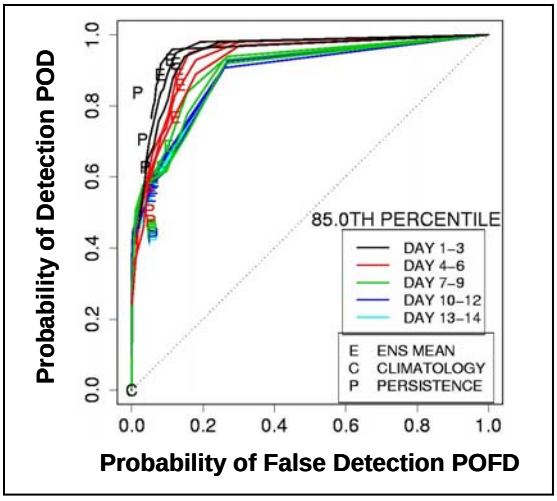
Plot of Reliability Diagram along with the histogram of sample sizes for a given event and for various lead times: EVS example. The histogram of sample sizes is given as Log(n) for all the

forecast probability bins and represents the forecast sharpness. In this case, the event is defined as being above the 75th percentile of the observed distribution. These results show that these ensemble forecasts are underspread or 'overconfident', particularly at short lead times; this is also reflected in the relative sharpness of the forecast probabilities (the forecasts are sharper for shorter lead time).



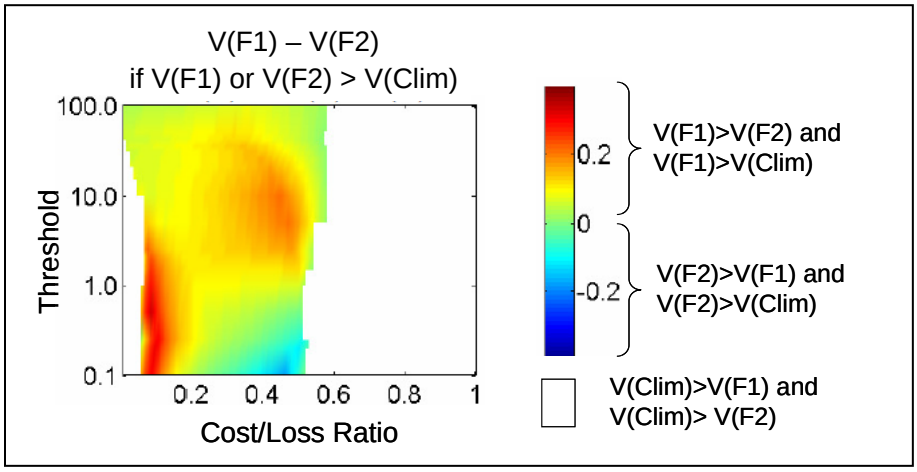
Example of Reliability Diagram plot for a given event and for various lead times

Plot of ROC Diagram for one specific event for single-valued forecasts and ensemble forecasts for several lead days: EVP example. In this case, the user can overlay (POFD, POD) values for single-valued forecasts (in this example given with points depicted with the letter) and probabilistic forecasts, using different colors for different lead times. The event is defined as being above the 85th percentile of the observed distribution in this example. By inter-comparing the ROC curves (and the ROC areas below the curves), these results show that the ensemble forecasts have more discrimination skill than any of the three single-valued forecasts for all lead times.



Example of ROC plot for both single-valued and probabilistic forecasts for a given event

2-D plot to compare Relative Value for different thresholds and different sets of forecasts: example from Environment Canada. Each event defined by one specific threshold leads to one Relative Value curve for a varying Cost/Loss ratio on the x axis. If the event thresholds are represented on the y axis, the values of the Relative Value score could be represented as a grid with colors. Here the user compares the Relative Value scores for two different forecast systems (climatology is the reference: when Relative Value <= 0, climatology is a better forecast).



Example of Relative Value plot for two sets of forecasts (climatology is the reference forecast for this skill score) as a function of threshold values

This section included examples of standard verification products for the four levels of information to initiate the discussion with the RFC forecasters and external users, about which verification products are the most meaningful. These proposed standards will have to be further evaluated by all the RFCs in collaboration with forecast users.

Enhancements to current verification applications and services

While working on the IVP and EVS exercises and the verification case studies, the RFC forecasters identified some required enhancements to the current verification applications and supporting applications. The list of IVP and EVS enhancements will help develop a unified verification system for CHPS-VS, which will combine the IVP and EVS functionalities to verify any single-valued and probabilistic forecasts. The graphical capabilities for this unified verification system will also use functionalities developed for the Graphics Generator within CHPS (especially for the plots of forecast and observed values, which are part of the Graphics Generator requirements).

Enhancements for IVP ob8.3

Here is the list of the enhancements relative to the IVP ob8.3 operational software, starting with the most critical ones.

1) ROC computation: modify ROC to compute one single POD value and one single POFD value for the user selected threshold; represent ROC curve as the curve from (0,0) to (POFD,POD), and to (1,1).
2) Scatter plots: include scatter plots of forecast error (forecast – observation) as a function of the observed value (to detect any conditional bias) and as a function of time.
3) Time series and scatter plots: allow the IVP batch process to generate these plots (currently these plots are generated only in the IVP GUI).
4) Temperature persistence: redefine persistence forecasts for temperature to maintain a diurnal cycle; the temperature persistent forecast should reproduce the last four 6-hourly observed temperature values for each day in the future (and not the last temperature observation for all 6-hourly time steps).
5) Metrics to add/modify: MAE-SS (IVP has only RMSE-SS), Relative Bias, ROC Score
6) Time series plots: make the observed time series more visible (the current color is light grey and the points are very small), with a line connecting the individual points similarly to the forecast time series; add a time series plot with the forecast error (forecast – observation) time series (currently only forecast and observed time series are plotted).
7) User control for x axis ordinates: the user should be able to re-order the x axis ordinates to provide more meaningful plots (e.g., re-ordering by locations).
8) Pop-up message: when the user is doing some data pooling across lead times, a window should pop up to recommend analyzing first the verification statistics at individual lead time in order to check whether data pooling across lead times would be meaningful.
9) Time aggregation functionality: add a functionality to verify forecasts at different time aggregated scales using a few statistics (minimum, maximum, average, and total), to allow users to verify multiple forecast products (e.g., minimum weekly flows derived from 6-hour instantaneous flows).
10) Selection of forecasts: allow the user to select valid time or basis time in IVP GUI.
11) Observations for pairing process: there is a pre-defined list of SHEF qualifier codes which make observations eligible for pairing. In some cases, the user might need to use observations with additional qualifier codes. Therefore the eligible qualifier codes should be specified in the pairing input file.

12) Pairing process: the pairing window has a lower limit of 1 hour and needs to accept floating point numbers (e.g., 0.25 for 15 minutes) to verify quickly changing conditions (i.e., tidal influences).
13) Current performance issues (speed and memory) when analyzing large datasets: even with certain "workarounds" to access more system memory and temporarily suspend posting to the RAX, some performance issues still persist and should be addressed in the future.
14) Batch mode run: provide an operational IVP batch menu to define all input files to run IVP in batch mode while reducing potential editing errors and improving operational efficiency.
15) Verification results for CHPS: enhance the IVP functionality to save the verification results as time series corresponding to statistics for multiple lead times, multiple time periods (e.g., months), or multiple event thresholds; these time series need to be imported in CHPS to be plotted in summary verification maps.

The most critical enhancements for IVP (from 1 to 6) should be taken care of in FY10 since the RFCs use the IVP software for their routine stage verification and will use it for their verification case studies.

Enhancements for EVS prototype version 1.0

Here is the list of the enhancements for the EVS prototype version 1.0, which was delivered to all the RFCs in May 2008. Some of these enhancements are already included in the EVS software version 2.0, which will be delivered to all the RFCs in October 2009. A detailed description of the software is provided in the EVS user's manual and in the EVS paper by Brown et al. (2009).

1) Sample size plots (done in EVS 2.0): add sample size plots, especially when using subsets of forecast-observed pairs.
2) Include skill calculations for CRPS and Brier Score using a given reference forecast (done in EVS 2.0).
3) Improve the algorithms for aggregating verification metrics across multiple forecast points (done in EVS 2.0).
4) Time series plots: add a time series plot for the user to display the ensemble values as box-and-whiskers (or only whiskers to improve plot readability) for all lead times, with the corresponding observations overlaid (there is already in EVS this time series plot for the forecast error (i.e., ensemble member – observation)). Time on the x axis should be given as dates (currently it is given as hours from the start date).
5) Scatter box-whisker plots: add other data display plots as proposed in the previous section (see examples from Allen Bradley).
6) Metrics to add/modify: add ROC score and CRPS decomposition (done in EVS 2.0); add Relative Value.
7) Include confidence interval computation for verification metrics and basic graphic capability.
8) User-friendliness: include more user friendly error messages to better explain what went wrong (done in EVS 2.0).
9) EVS User Manual (done in EVS 2.0): add easy to understand and real examples of graphics from EVS, as well as labels to help users interpret the results; these labels should be available in the EVS help along with the current links and mathematical formulas.
10) Allow transformations between imperial and metric units for the most common measurement units (e.g., CFS to CMS or INCH to MM and vice versa) (done in EVS 2.0).

11) Include R-scripts to produce high quality EPS graphics from the numeric outputs from the EVS for scientific manuscripts and reports (done in EVS 2.0).
12) Verification results for CHPS: enhance the EVS functionality to save the verification results as time series corresponding to statistics for multiple lead times, multiple time periods (e.g., months), or multiple event thresholds; these time series need to be imported in CHPS to be plotted in summary verification maps.

The RFCs will continue to evaluate the EVS version 2.0 to develop additional requirements for the EVS and for the unified verification system for CHPS-VS.

Scientific enhancements

One of the most needed enhancements for flow forecast verification is the characterization of the timing error, the magnitude error, and the hydrograph shape error. The estimation of the timing error could be very valuable for forecast users (e.g., navigation industry using tidal forecasts), and to further diagnose the main error sources in the forecasts. Also, as NCRFC showed in their verification case study (see Appendix A), the errors in the rising limb and in the falling limb could be significantly different. The difficulty is to define an observed event and a forecast event to be paired together (in the current verification process, pairing is based on forecast and observed valid time). The RFCs recommended developing first a simple manual pairing process (similar to the pairing process available in the STAT-Q tool for calibration purposes), and in the future a combination of automated and manual event pairing (since a fully automated process may lead to incorrect pairing when the events are quite complex). OHD is already working on this enhancement, including the use of wavelet analysis and how existing spatial verification methods and curve registration techniques can be adapted for error analysis of flow time series.

Another important scientific enhancement currently being developed by OHD is a facility to compute and display measures of sampling uncertainty, such as confidence intervals, for all of the verification metrics. Verification is unavoidably based on datasets of finite samples that do not fully represent the true underlying distribution of the entire population, leading to verification statistics that are prone to sampling errors. In other words, sample size is an important concern when verifying forecast probabilities, especially for extreme events. Thus, an appreciation of sampling uncertainty is important when interpreting the verification results. Currently, the EVS presents plots of sample counts for each metric but does not include measures of sampling uncertainty of the metrics. Work is underway to derive (analytically and numerically, through bootstrapping technique) estimates of sampling uncertainty to be integrated into EVS.

Besides, to analyze the sources of forecast uncertainty, all data inputs used and all output produced by the river forecast system must be verified. Therefore forecast verification needs to be applied and tracked across the entire NWS forecast process. Weather, climate, and water forecasts need to be evaluated using verification metrics and parameters of hydrological relevance. This requires close collaborations between the weather community and the hydrologic community to use verification measures and practices appropriately for hydrological applications. Work is underway with NCEP/EMC to provide more consistent verification information for ensemble forecasts, in particular for the verification products available at http://www.emc.ncep.noaa.gov/gmb/yzhu/ and which could also be provided for the NAEFS ensemble verification. This includes using the same verification metrics (e.g., CRPS, Reliability

Diagram, ROC, Relative Value) and verifying ensembles at the same spatial and temporal scales (e.g., verification statistics for daily forecasts for each RFC area). In particular, OHD and NCEP agreed to use a set of RFC-defined masks to present NCEP grid verification statistics on various spatial areas within the RFC areas (e.g., RFC area, carryover groups, and forecast groups). This would be especially useful to the RFCs that are directly ingesting NCEP ensemble forcing inputs to produce hydrologic ensemble forecasts since these ensemble inputs have some known bias in the ensemble mean and the ensemble spread.

Enhancements of supporting applications

Storage requirements of verification data is significant since all forcing inputs and hydrologic outputs, including observations, simulations, forecasts and hindcasts for potentially different forecasting scenarios, as well as metadata (e.g., description of forecasting scenario), forecast point attributes (e.g., impact thresholds), and verification statistics must be retained for statistical analysis. Therefore a key component to support a comprehensive verification service is a robust archive system with back up, data viewing and quality control functionality. Even if the hindcast data do not necessarily need to be archived if one could regenerate them in a reasonable time period, archiving all single-valued and probabilistic operational forecasts for all forecast points requires a very significant effort in terms of system design, hardware and software.

The NWS Verification Team provided a report on data archiving requirements for forecasting and verification to the IWT Archive Team led by Julie Meyer; these requirements are available in the interim team report at
http://www.nws.noaa.gov/oh/rfcdev/docs/NWS-Verification-Team_interim_report_Jan09.pdf

The second supporting application is the hindcasting capability to retroactively generate forecasts using a fixed forecasting scenario for a large time period (whereas the operational forecasts produced in real time could come from a forecasting system changing in time). This requirement is especially needed for probabilistic forecasts. Several years of forecasts/hindcasts are needed to verify forecast probabilities since it is impossible to determine whether a single probabilistic forecast is correct or incorrect based on a single outcome. For any type of forecasts (single-valued and probabilistic), the verification of extreme events requires a long archive of forecasts/hindcasts to get large enough sample size of these extreme events and produce reliable verification metrics.

While hindcasting capabilities in NWSRFS were very limited, the CHPS prototype using the FEWS core capabilities offers hindcasting capabilities for both single-valued and probabilistic forecasts. Initial experiments of ensemble hindcasting using the CHPS prototype in FY09 have indicated that this hindcasting capability would meet all the foreseeable needs of the modelers and forecasters (the implementation of hindcasting in CHPS by the HSMB is currently under way for ensemble science evaluation purposes).

Hindcasting could also be useful for post-event analysis. Forecasters may investigate how the forcing inputs, model states and model parameters impact the forecast made for a specific event by running different forecasting scenarios (see in Appendix A the NWRFC verification case study using IFP to produce the different forecast scenarios). Once the main error sources have been identified for one specific event, the forecaster should identify similar historical events and

evaluate if the forecasts for these historical events have similar main error sources. This is related to the real-time verification approach: the user queries the database of archived forecasts to select a few analogous forecasts and their corresponding observations and identify potential errors in the real-time forecast, based on errors identified in the historical analogs. Such functionality for real-time verification is currently under development and will be made available in CHPS.

Verification training

Training on forecast verification is very much needed for many reasons: verification includes multiple metrics; some metrics are quite complex (Reliability Diagram for example); how to analyze forecast performance depends on many factors (basin characteristics, atmospheric and hydrologic conditions, space and time scales, etc.). Progress is being made on scientific enhancements and software development, which needs to be presented to the RFCs on a regular basis; their feedback helps develop verification software and scientific enhancements that meet their needs and the needs of their forecast users.

The first COMET training module on hydrologic forecast verification, which was delivered in June 08, was very well received by all the RFC team members. COMET is currently developing two other modules: a QPF verification module and a module on hydrologic verification case studies with IVP and EVS (based on expanding two RFC case studies).

The team proposed that training on IVP and EVS, such as the software demonstrations given at the two verification workshops and at the verification team meetings, should be recorded as webinar to be offered at any time. Go-To-Meetings should also be held regularly to answer specific questions on software and interpretation of verification graphics. The participants of the two RFC verification workshops in August 2007 and November 2008 recommended conducting regular verification workshops (the next workshop could be conducted in FY2011), with the verification team members and a few extra RFC participants, to share verification experiences and show progress being made in the NWS and in academia.

When verification products are made available to the public, self-learning modules should also be provided so that the users can better understand the verification metrics and plots and efficiently utilize the information for their decision making.

Future activities for the NWS Hydrologic Forecast Verification Team

The verification team has made a lot of progress in the last two years. As a result, verification case studies are set up at all RFCs and the RFC feedback has been integrated in the OHD research and development verification activities. This report proposes standard verification strategies and products for five overarching questions. The verification team has identified further work needed to produce, evaluate, and improve these standards and help develop a comprehensive CHPS verification service that meets all user needs. These near-term activities are outlined below.

Evaluate proposed verification standards with RFC case studies

Because forecast quality is multi-faceted and verification can be performed for various purposes, the verification standards proposed in this report include multiple metrics and products, as well as analyses of different forecasting scenarios. The RFCs need to produce and evaluate these verification standards by working with their own datasets to provide feedback and help improve the standards. What-if scenarios need to be identified to specify which datasets, which spatial and temporal scales, which data stratification, and which verification metrics and products should be used to perform meaningful verification analyses for a range of situations (e.g., drought forecasts, flood forecasts, record forecasts, and tidal forecasts) and for a range of applications. By carrying out verification case studies, each RFC should determine which verification products are the most meaningful for their forecast users and demonstrate how verification helps guide improvements of the forecasting system in their RFC area. Also the verification studies at all RFCs will help identify the best ways to aggregate verification summary statistics across multiple forecast points for forecast performance tracking (see section on questions 4 and 5). The team should identify criteria for deriving meaningful subsets of forecast points on which summary verification scores can be reported at the national level. The team should develop prototype capabilities to produce the standard verification products using IVP, EVS, and the CHPS display capability, which are the Graphics Generator, the FEWS Spatial Display, and the FEWS Time Series Display).

The team has identified two analyses that should be performed at all RFCs (see section on question 3), namely the impact of QPF forecast horizon on hydrologic forecast quality and the impact of run-time modifications performed on the fly by the forecasters. Other RFC case studies could be set up, either to expand the case studies presented in the interim team report (e.g., recent forecast data could significantly increase the sample sizes and provide more robust verification results) or tackle new forecast performance issues, as described in this report. Working with CHPS is likely to facilitate the definition of the various forecasting scenarios to be run in parallel in order to inter-compare verification results. Therefore it is expected that the CAT RFCs will first implement these forecasting scenarios in CHPS, while the CAT-II RFCs work on their CHPS migration first and, once it is completed, on the implementation of the different forecasting scenarios.

Support the CHPS Verification Service development

OHD is currently developing in collaboration with Deltares prototypes for CHPS-VS for both diagnostic verification and real-time verification. RFC forecaster feedback is essential in the design and prototyping phase since the main goal of the verification system is to help forecasters improve forecasts. The verification team should support the design and development of CHPS-VS by providing input and reviewing user requirements and software design documents, and by evaluating prototype functionality (e.g., future version of EVS).

Perform detailed user analysis of verification products

A few groups of users have been identified in this report to propose four levels of sophistication when providing verification products to these users. Further user analysis will be required with the RFC Service Coordination Hydrologists and OCWWS to better understand which verification products are the most meaningful for each user group. This effort should also be coordinated with the verification efforts from the meteorological community to present consistent verification information for weather forecasts, climate forecasts and hydrological forecasts.

Define requirements for disseminating verification information to users

Current verification statistics compiled by the NWS Performance Branch (called stats-on-demand) aggregate forecasts over time to compute basic error statistics for individual forecast points. These statistics are then averaged over various response times and geographical extents. Because there is no information to place these statistics in context or information to distinguish individual events, this verification system has been of little value. The NWS Performance Branch is aware of the recent progress made at OHD and the RFCs on hydrologic forecast verification since Julie Demargne is a member of the National Performance Branch Committee (NPMC).

As proposed in the FY09 verification work plan, the team should develop requirements to improve the routine hydrology verification statistics computed and archived by the NWS Performance Branch. The team should define which standard verification products should be disseminated to the users by the NWS Performance Branch and by the RFCs. These requirements should be based on results from the case studies completed and ongoing, as well as the standard verification strategies described in this report. These requirements should be presented to the RFCs and the OHD management. The NWS Performance Branch should be engaged as soon as possible so that these requirements can become part of their work plan.

In order for the team to continue its work on verification, a second team charter has been developed and is given in Appendix C. This team charter has been presented at the HIC meeting on July 10, 2009 and has been reviewed by the HICs. This team charter will be finalized in October 2009 to make sure that the verification team has identified meaningful future activities with reasonable deliverables, resources and schedule.

Conclusions and recommendations

This team report proposes standard strategies and products for hydrologic forecast verification to answer the following five overarching questions:

1) How good are the forecasts?

2) What are the strengths and weaknesses of the forecasts?

3) What are the sources of uncertainty and error in the forecasts?

4) How are new science and technology improving the forecasts?

5) What should be done to improve the forecasts?

Given the variety of forecast applications and the different attributes of forecast quality, four different levels of verification information, each containing several key verification metrics and products, have been identified to meet the needs of all users. These levels of verification information should be provided for the verification analyses relative to the five overarching questions, although the first three levels (data display plots and verification scores on individual forecast points and on spatial maps) should be sufficient for most users.

Recommendations for the verification analyses include:

- The impact of any newly developed forecast process (e.g., new calibration parameters, new preprocessing technique, new observed dataset) should be analyzed via systematic verification. When comparing two forecasting scenarios, the verification results need to be produced for the exact same events (e.g., same verification period, same time step). Therefore verification results should be reported separately for the daily forecast points and the flood only points.
- For forecast performance tracking purposes, it is necessary to define meaningful groups of basins to aggregate the verification summary scores (skill scores, as well as reliability, resolution and discrimination summary measures) without masking potential forecast improvement. By working on verification studies at all RFCs, the verification team plans to propose criteria (such as basin response time) to define subsets of forecast points with similar hydrologic processes.
- The use of normalized metrics (e.g., skill scores) and metrics defined for common probability thresholds (e.g., from the observed probability distribution), rather than absolute thresholds for basins with different flow characteristics, is necessary when comparing verification results across different basins and aggregating these results (if the basins show similar verification characteristics).
- Temporal aggregation is necessary to verify different forecast products (6-hourly instantaneous flow forecasts vs. weekly minimum flow forecasts) and the verification team should define a few time scales for forecast verification to support specific users. Besides forecasts should be verified first for each individual lead time since forecast performance varies greatly with lead time. If the verification statistics show similar

characteristic across multiple lead times, one could then pool the forecast data from this subset of lead times to increase the sample size.

- Spatial aggregation of verification results across different basins should be carefully performed, not to mask large variations of forecast performance among the basins. Verification statistics should be first analyzed for individual basins and plotted on spatial maps to define subsets of basins for which the verification results have similar characteristics.
- Forecast performance should be evaluated different conditions by stratifying the forecast-observed dataset based on both time conditioning (e.g., by season and by month) and atmospheric/hydrologic conditioning. For inter-comparison purposes, the verification team should agree on a few categories for data stratification, using low and high thresholds defined from the observed probability distribution, and using specific absolute thresholds (e.g., probability of precipitation, freezing level, or flooding level). It is important to not define too many categories so that the sample size for each category contains enough data to give reliable verification statistics.
- Verification results should be reported along with the sample sizes since the sampling uncertainty could have a significant impact on the values of the verification statistics for small sample sizes (which is usually the case for extreme events). Work is underway to estimate and represent the sampling uncertainty in the verification metrics with confidence intervals. Once the verification software has the capability to estimate confidence intervals, the verification measures should be accompanied by the confidence intervals for a given confidence level.
- The different sources of uncertainty and error need to be analyzed by verifying both the forcing input forecasts and the hydrologic outputs. For extreme events, both flow forecasts and stage forecasts should be verified since verification results of flow and stage could be significantly different due to the quality of the rating curves for such events. Sensitivity analysis of the different sources of uncertainty relies on using different forecasting scenarios. Two sensitivity analyses for single-valued stage forecasts are recommended for all RFCs: 1) impact of the QPF horizon on the hydrologic forecast performance, by using QPF forecasts of increasing horizons (from 6-hourly to 5 days); 2) impact (on a day-to-day basis) of run-time modifications made on the fly on the hydrologic forecast performance, by using two forecasting scenarios with and without run-time modifications made on the fly (but including the a priori modifications, which are defined by the forecasters before running any forecast). The verification team recommends a set of common baseline scenarios to be used at all RFCs, although each RFC could define additional scenarios to meet specific local needs (e.g., stage forecasts produced from longer QPF horizon).

Required enhancements of the current IVP and EVS software and verification science are identified (some of the proposed scientific enhancements were already included in the OHD verification activities for FY09). The main enhancements concern: 1) the analysis of timing error information of flow forecasts; 2) the estimation of confidence intervals (along with a graphical capability) to represent the sampling uncertainty of the verification metrics; 3) the consistency of verification information for weather and water forecasts; work is underway with NCEP to use similar verification metrics and report results for spatial areas that are consistent with the

hydrological modeling performed by the RFCs. (e.g., verification statistics for each RFC area). Additional efforts are needed for data archiving (which is crucial to archive all data and metadata required for verification), hindcasting (to retroactively generate forecasts from a given scenario with large enough sample size), as well as verification training for RFC forecasters and forecast users.

Finally future activities for the verification team are proposed to:

- Produce, evaluate and improve the verification standards with expanded verification case studies at all RFCs. In their verification case studies, the RFCs should determine which verification products would be the most meaningful for them and for their forecast users. They should also demonstrate how verification would help guide improvement of the forecasting system and the forecast process. The team will develop prototype functionalities to produce the verification standards with the existing software (IVP, EVS, WR water supply website, and the CHPS display capabilities). Such analysis will help define criteria to aggregate verification results across basins and track forecast performance on the identified groups of basins.
- Define what-if scenarios to specify which observed and forecast datasets, spatial and temporal scales, verification metrics and products should be used for a range of situations (e.g., drought forecasts, flood forecasts, record forecasts, and tidal forecasts) and for a range of applications.
- Perform detailed user analysis of the verification products in collaboration with the RFC Service Coordination Hydrologists and OCWWS and develop requirements for dissemination of verification information for the RFC river forecasts by the NWS Performance Branch and by the RFCs (the verification products accompanying the forecast products).
- Continue to support the design and development of the CHPS Verification Service (CHPS-VS) by testing verification prototypes (e.g., EVS) and reporting requirements and necessary enhancements for a unified verification system that meets all user needs.

A second team charter is proposed in Appendix C to perform these future activities from October 2009 to September 2011.

The proposed verification standards are likely to evolve as new verification science and software are being developed, for example to account for the uncertainty in the observations, to verify extreme events and account for climate change, and to verify spatial and temporal joint distributions (not only forecasts at a single location for one specific lead time). Collaborative research work is under way with universities (e.g., University of Iowa, University of California, Irvine, Iowa State) and NCEP/EMC, as well as scientists involved in the Hydrologic Ensemble Prediction Experiment (HEPEX) verification test-bed (which involves Environment Canada and ECMWF). The role of the verification team (which includes all RFCs) seems essential to ensure that these collaborative efforts will lead to common verification products and practices for weather, climate and water forecasts, thus meeting the needs of all forecast users.

References

Brown J.D., Demargne J., Seo D-J., Liu Y., 2009. The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Submitted to *Environmental Modelling and Software*.

Demargne J., Mullusky M., Werner K., Adams T., Lindsey S., Schwein N., Marosi W., Welles E., 2009: Application of Forecast Verification Science to Operational River Forecasting in the U.S. National Weather Service. *Bulletin of the American Meteorological Society* 90 (6): 779-784.

Jolliffe I.T. and Stephenson D. B., 2003: Forecast Verification, A Practitioners Guide in Atmospheric Sciences, Wiley, West Sussex, England, 240 pp.

Taylor K.E., 2001: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research 106 (D7), 7183–7192.

Welles E., 2005: Verification of River Stage Forecasts, Dissertation, University of Arizona, 155 pp.

Welles E. and Sorooshian S., 2009: Scientific Verification of Deterministic River Stage Forecasts, Journal of Hydrometeorology, 10 (2), 507-520.

Wilks D.S., 2006: Statistical Methods in Atmospheric Sciences, Academic Press, San Diego, California, 627 pp.

World Meteorological Organization (WMO), 2004: WWRP/WGNE Joint Working Group on Verification, Forecast Verification – Issues, Methods and FAQ, web site: http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

World Meteorological Organization (WMO), 2008: Recommendations for the Verification and Intercomparison of QPFs and PQPFs from Operational NWP Models, WWRP 2009-1, 37 pp.

Appendix A – Description of the verification case studies from the 13 RFCs March-November 2008

As part of the NWS Hydrologic Forecast Verification Team, the 13 RFCs worked on a verification case study using one of the two available verification applications: the IVP operational software for single-valued forecast verification, and the EVS prototype software for ensemble forecast verification. These studies included single event analysis and multiple year forecast analysis, using the operational forecasts as well as experimental forecasts from different scenarios. The goals of these case studies were to build some verification expertise at the field offices, to test the available software to identify their limitations, and to analyze how verification could help the NWS improve their forecasts by doing such post-event and diagnostic verification analysis.

The case studies were presented by the 13 RFCs in the interim team report available at http://www.nws.noaa.gov/oh/rfcdev/docs/NWS-Verification-Team_interim_report_Jan09.pdf Here is a short description of the case studies.

- CNRFC: post-event analysis of operational forecasts of precipitation, temperature and flow during a major flood event and for 2 river basins
- NWRFC: post-event analysis and retrospective analysis of the error sources (using IFP) for one extreme event and for one basin
- CBRFC: analysis of the impact of QPF errors on flow forecasts using raw model forecasts and flow simulations for a flood event and for 2 basins
- LMRFC: analysis for forecast performance during Hurricane Katrina in comparison to forecasts from the last five years, for 4 forecast points
- NCRFC: performance analysis of operational forecasts and QPF contingency forecasts during the record flooding events in 2008 for 4 forecast points
- MBRFC: performance analysis of operational forecasts of flow and stage for four basins, with persistence as the baseline forecast
- SERFC: performance analysis of operational forecasts for one basin during the last 8 years and proposed analysis of errors in forecast shape and forecast timing
- NERFC: performance analysis of QPF forecasts from various sources (HPC, NDFD, and NERFC) and produced by individual forecasters on 13 basins
- OHRFC: uncertainty analysis for operational forecasts with 6 sets of experimental forecasts from different scenarios (MODs/no-MODs, no-QPF/HPC-QPF/HAS-QPF) for 7 basins
- APRFC: impact of two calibration strategies on stage forecast quality for one basin
- WGRFC: impact of VAR state updating procedure on stage forecast quality by comparing VAR forecasts with operational forecasts on 3 forecast points

- ABRFC: performance analysis of ensemble hindcasts produced by the HMOS prototype software for 3 basins
- MARFC: performance analysis of ensemble forecasts produced by the Ensemble Preprocessor prototype (EPP2) and ESP for 2 basins

The presentations of these verification case studies are also available online: for NW-, SE-, AP-, and NC-RFCs at
http://www.nws.noaa.gov/oh/rfcdev/projects/rfcHVT_workshop2_agenda_presentations.html
and for the other RFCs at
http://www.nws.noaa.gov/oh/rfcdev/projects/rfcHVT_mtg_docs.html

Appendix B – Glossary of verification metrics

Bias
The difference between the mean of the forecasts and the mean of the observations. Could be expressed as a percentage of the mean observation. Also known as overall bias, systematic bias, or unconditional bias.
Relative Bias is computed as: RB = Mean Error / (observed mean).
Another relative measure is the Percent Bias:

$$PB = 100 \times \left[\sum_{i=1}^{n} \left(Fcst_i - Obs_i \right) / \sum_{i=1}^{n} \left(Obs_i \right) \right]$$

For categorical forecasts, bias (also known as frequency bias) is equal to the total number of events forecast divided by the total number of events observed. With the (2x2) **contingency table**, Bias = (a+b)/(a+c). Perfect score: 1.

Brier Score (BS)
The mean square error of probabilistic two-category forecasts where the observations are either 0 (no occurrence) or 1 (occurrence) and forecast probability may be arbitrarily distributed between occurrence and non-occurrence. BS=0 for perfect (single-valued) forecasts. BS=1 for forecasts that are always incorrect.

Brier Skill Score (BSS)
A **Skill Score** based on **BS** values. The most commonly used reference forecasts are persistence and climatology. Perfect score: 1.

Contingency Table
A two-dimensional table that gives the discrete joint distribution of forecasts and observations in terms of cell counts. For dichotomous categorical forecasts, having only two possible outcomes (Yes or No), the following (2x2) contingency table can be defined:

2x2 Contingency Table		Event Observed	
		Yes	**No**
Event Forecast	**Yes**	**a** (hits/true positives)	**b** (false alarms/false positives)
	No	**c** (misses/false negatives)	**d** (true negatives)

Continuous Ranked Probability Score (CRPS)
A measure of the integrated squared difference between the cumulative distribution function of the forecasts and the corresponding cumulative distribution function of the observations. It is an extension of the **Ranked Probability Score (RPS)** for continuous probability forecasts. It corresponds to the **Mean Absolute Error (MAE)** for single-valued forecasts. Perfect score: 0.

Continuous Ranked Probability Skill Score (CRPSS)
A **Skill Score** based on **CRPS** values. The most commonly used reference forecasts are persistence and climatology. Perfect score: 1.

Correlation Coefficient
A measure of the linear association between forecasts and observations independent of the mean and variance of the marginal distributions. Pearson Correlation Coefficient and Spearman Rank Correlation are the most widely used ones. Perfect score: 1.

Discrimination Diagram
A diagram plotting the conditional distributions of the forecasts. For binary events, this diagram plots the conditional distribution of the forecasts given that the event occurred, and the conditional distribution of the forecasts given that the even did not occur. Ideally, the two distributions are well separated from one another, becoming two distinct spikes for perfect forecasts.

False Alarm Ratio (FAR)
For categorical forecast, the number of false alarms divided by the total number of events forecast. A measure of reliability. With the (2x2) **contingency table**, FAR = b/(a+b). Not to be confused with the **Probability of False Detection (POFD)** (also called **False Alarm Rate**) (which is conditioned on observations rather than forecasts). Range: 0 to 1. Perfect score: 1.

Lead Time of Detection (LTD)
The average lead time of forecasts that correspond to hits in the contingency table.

Mean Absolute Error (MAE)
The average of the absolute differences between forecasts and observations. A more robust measure of forecast accuracy than Mean Square Error that is sensitive to large outlier forecast errors. It corresponds to the **Continuous Ranked Probability Score (CRPS)** for probabilistic forecasts. Perfect score: 0. Note: the overbar denotes the mean.

$$MAE = \overline{(|\ f - o\ |)}$$

Mean Absolute Error Skill Score (MAE-SS)
A **Skill Score** based on **MAE** values. The most commonly used reference forecasts are persistence and climatology. Perfect score: 1.

Mean Error (ME)
The average difference between forecasts and observations. Note: it is possible to get a perfect score if there are compensating errors. Perfect score: 0.

$$ME = \overline{(f - o)}$$

Probability Of Detection (POD) (or Hit Rate)
For categorical forecast, the number of hits divided by the total number of events observed. A measure of discrimination. For the (2x2) **contingency table**, POD = a/(a+c). Range: 0 to 1. Perfect score: 1.

Probability Of False Detection (POFD) (or False Alarm Rate)
For categorical forecast, the number of false alarms divided by the total number of events observed. A measure of discrimination. For the (2x2) **contingency table**, POFD = b/(b+d). Not

to be confused with the **False Alarm Ratio (FAR)** (which is conditioned on forecasts rather than observations). Range: 0 to 1. Perfect score: 0.

Root Mean Square Error (RMSE)
The square root of the average of the squared differences between forecasts and observations. It puts a greater influence on large errors than smaller errors, which may be good if large errors are especially undesirable, but may also encourage conservative forecasting. Perfect score: 0.

$$RMSE = \sqrt{\overline{(f-o)^2}}$$

Ranked Probability Score (RPS)
The mean square error of probabilistic multi-category forecasts where observations are 1 (occurrence) for the observed category and 0 for all other categories and forecast probability may be arbitrarily distributed between all categories. By using cumulative probabilities, it takes into account the ordering of the categories. For two category forecasts, the RPS is the same as **Brier Score**. Perfect score: 0.

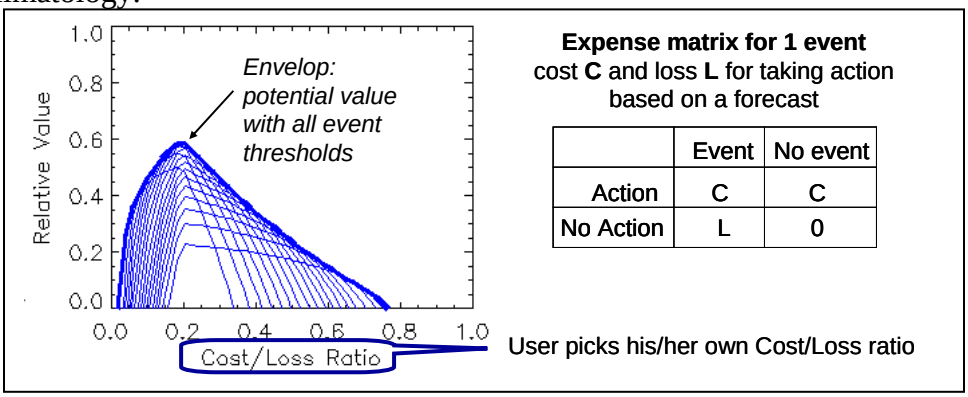
Ranked Probability Skill Score (RPSS)
A **Skill Score** based on **RPS** values. The most commonly used reference forecasts are persistence and climatology. Perfect score: 1.

Relative (or Receiver) Operating Characteristic (ROC)
A signal detection curve for binary forecasts obtained by plotting **POD** (y-axis) versus **POFD** (x-axis) to describe the forecast discrimination. There is one curve for each set of forecast-observed pairs and for a given event. The ROC curve for single-valued forecasts is defined as the curve from (0,0) to (POFD, POD) and to (1,1). For probabilistic forecasts, there are N increasing probability levels (binary classifiers) to turn the probabilistic forecast into a yes/no forecast; the ROC curve for probabilistic forecasts is defined as the curve from (0,0) to ($POFD_k$, POD_k) for each k^{th} probability level from 1 to N, and to finally (1,1). The ROC curves for single-valued forecasts and probabilistic forecasts for a given event can directly be inter-compared if plotted together. The 45 degree diagonal line indicates no skill. It is conditioned on the observations (given that Y occurred, what was the corresponding forecast?). It is a good companion to the **Reliability Diagram**, which is conditioned on the forecasts. Perfect: curve travels from bottom left to top left of the diagram, then across to top right of the diagram.

ROC Score
A summary score for binary forecasts derived from the **ROC** curve and its ROC Area (area below the ROC curve) for a given event to describe the forecast discrimination.
ROC Score=2 x (ROC Area – 0.5). Perfect score: 1.

Relative Value (or Economic Value)
Skill score of expected expense using a Cost/Loss ratio, with climatology as a reference. The 2x2 expense matrix is defined for a given event with cost C for taking action based on a forecast (the event being observed or not) and loss L for taking no action when the event actually occurred. The expense matrix is multiplied by the 2x2 contingency table to estimate the expense for the specified event. Perfect score: 1.

Since the Relative Value depends on the Cost/Loss ratio, it is plotted as a curve for Cost/Loss ratio varying from 0 to 1 for a given event. When considering a range of events, all the Relative Value curves are plotted together and the envelop of all the curves represents the potential economic value. For probabilistic forecasts, one needs to produce a curve for each probability threshold at which the forecast says the event will occur (similarly to the ROC curve with one point for each probability threshold). As for any skill score, if Relative Value is greater than zero, the forecast has more potential value than climatology; otherwise the forecast is worse than climatology.



Reliability Diagram

A diagram in which the frequency of the observations, given the forecast probability, is plotted against the forecast probability, where the range of forecast probabilities is divided in to K bins. The sample size in each bin is often included as a histogram or values beside the data points. Perfectly reliable forecasts have points that lie on the 45 degree diagonal line. The deviation from the diagonal line gives the conditional bias. The Reliability Diagram is called the Attributes Diagram when the no-resolution line and the no-skill line with reference to climatology are included. It is conditioned on the forecasts (given that X was predicted, what was the outcome?). It is a good partner of the **ROC**, which is conditioned on the observations.

Root Mean Square Error Skill Score (RMSE-SS)

A **Skill Score** based on **RMSE** values. The most commonly used reference forecasts are persistence and climatology.

Sample Size

A numeration of the number of forecasts involved in the calculation of a metric appropriate to the type of forecast (e.g., categorical forecasts should numerate forecasts and observations by categories, etc.)

Skill Score

A measure of the relative improvement of the forecast over some (usually 'low-skilled') benchmark forecast. Skill score is associated with a given verification metric and a given reference forecast. Commonly used reference forecasts include climatology, persistence, or output from an earlier version of the forecasting system. Perfect score: 1.

$$SS = \frac{Score(\,forecast) - Score(reference)}{Score(\,perfect) - Score(reference)}$$

Note: if the score of perfect forecast is equal to 0 (e.g., for MAE and CRPS), the skill score is computed as:

$$SS = 1 - \frac{Score(\ forecast)}{Score(reference)}$$

Talagrand Diagram (or Rank Histogram)
A plot of observed frequencies for k non-overlapping bins of equal probability for the forecast distribution. It measures how well the observed probability distribution is represented by the forecasts. For perfect forecasts, the rank histogram is flat since the observation is equally likely to fall between any two members. For U-shaped histogram, the ensemble spread is too small, most observations falling outside the extremes of the ensemble. For dome-shaped histogram, the ensemble spread is too large, most observations falling near the center of the ensemble. For asymmetric histogram, the model has a bias to one side.

Uncertainty
The degree of variability in the observations. Most simply measured by the variance of the observations. Important aspect in the performance of a forecasting system, over which the forecaster has no control.

On-line References
http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html
http://www.swpc.noaa.gov/forecast_verification/Glossary.html

Appendix C – Proposed second team charter
September 30, 2009

Vision: River forecast verification tools will be readily available to various users including forecasters, service hydrologists, managers, and the general public to produce verification information that is meaningful to each user group. Standard strategies and products for hydrologic forecast verification will help answer the following questions: 1) How good are the forecasts? 2) What are the strengths and weaknesses of the forecasts? 3) What are the sources of uncertainty and error in the forecasts? 4) How are new science and technology improving the forecasts? 5) What should be done to improve the forecasts? RFC forecasters and modelers will systematically generate and communicate river forecast verification results and identify, based on routine verification, how the forecasting system may be improved to meet the evolving service needs. Ultimately, forecast verification is considered successful when its results are used by forecasters and modelers to guide improvement of the forecasting system and the forecast process, and by users to maximize the utility of forecast information in their decisions.

Statement of the Problem: Currently, information on NWS river forecast performance is limited in scope and generally not communicated to most user groups. In recent years, nationally-supported verification software has been developed, which will be integrated in the near future into the CHPS Verification Service (CHPS-VS). Also verification case studies have been carried out at all RFCs and recommendations on standard verification metrics and products, as well as verification analyses, have been developed by this team as described in the September 09 team report. For verification to become a routine operation at the RFC, these verification standards need to be field-tested and further evaluated for improvement by the RFCs with verification case studies. Also, in order for CHPS-VS to meet all user needs, more detailed analysis of user requirements is needed.

Mission: Carry out the following to communicate meaningful river forecast verification information to user groups including forecast users, forecasters, service hydrologists, and managers using existing software (IVP and EVS) and to support the development of CHPS-VS:

(1) Generate and evaluate the standard verification metrics and products recommended in the September 09 team report with new RFC verification case studies;

(2) Perform detailed user analysis of the verification products in collaboration with the RFC Service Coordination Hydrologists and OCWWS and develop requirements for disseminating verification information for RFC river forecasts to users;

(3) Support the design and development of CHPS-VS by providing input and reviewing user requirements and software design documents and by testing prototype functionalities (e.g., EVS).

Success Criteria: The team will develop a verification report by September 30, 2011 that describes improved standard verification metrics and products, as well as RFC case studies using these standards. The RFC case studies will include the analyses of the impacts of the QPF horizon and the run-time modifications made on the fly on the quality of river stage forecasts. To

accomplish this, each RFC will develop a report on their verification case study to show the quality of the forecasts under different conditions and/or for different forecasting scenarios, and the value of verification for them and for their forecast users. The team leader will describe verification analyses of experimental ensemble forecasts for various RFC test basins to evaluate new science. All the verification case studies will include a discussion on how verification helps guide improvement of the forecasting system and the forecast process. The team will also coordinate with the RFC Service Coordination Hydrologists (SCHs) and OCWWS to define which standard verification products should be disseminated by the NWS Performance Branch and by the RFCs. The team will deliver the prototype capabilities developed at OHD and at the RFCs to produce the standard verification products using IVP, EVS, and the CHPS display capability (provided by the Graphics Generator, the FEWS Spatial Display, and the FEWS Time Series Display) and to disseminate the verification information along with forecast products to users.

Scope of Authority / Limitations: The team will normally meet every two/three months via teleconference. Team members will share their progress on the verification case studies and on the prototype capabilities to produce standard verification products using existing software. The team will also evaluate the recommended standard metrics and products and improve them if necessary. The team will collaborate with the RFC SCHs and OCWWS to perform a more detailed user analysis of the verification products needed for all user groups. The user analysis may require sub-groups of verification experts (including experts from the NWS Performance Branch and outside the NWS), OCWWS experts, and RFC SCHs to focus on specific forecast users. The team will develop requirements for disseminating verification information for RFC river forecasts to users. The team will also review user requirements and software design documents for the CHPS Verification Service and test prototype functionality to report required enhancements. A verification workshop will be organized in FY 2011 to share progress on verification science, software and case studies made in the NWS, other agencies, and academia. The team should realize the success criteria defined here no later than September 30, 2011.

No travel beyond the RFC verification workshop is authorized for these goals.

Proposed Team Membership:
Team will be comprised of verification focal points and co-focal points from the 13 RFCs.

Julie Demargne (OHD/HSMB) - lead
Ernie Wells (OCWWS/HSD)
Larry Lowe (ABRFC verification focal point)
James Coe (APRFC verification focal point)
Kevin Werner (CBRFC verification focal point and SCH coordinator)
Alan Takamoto (CNRFC verification focal point)
Kai Roth (LMRFC verification focal point)
Bill Marosi (MARFC verification focal point and NWSEO representative)
Andrew Philpott (MARFC verification co-focal point)
Julie Meyer (MBRFC verification focal point)
Holly Reckel (NCRFC verification focal point)
Tom Econopouly (NERFC verification focal point)

Steve King (NWRFC verification focal point)
Tom Adams (OHRFC verification focal point)
Christine McGehee (SERFC verification focal point)
Greg Waller (WGRFC verification focal point)

Verification software experts in OHD will serve as technical advisors to the team:
James Brown (OHD/HSMB)
Yuqiong Liu (OHD/HSMB)
Hank Herr (OHD/HSEB)

Proposed Schedule:
The schedule reflects the need for the RFCs to work on the CHPS implementation first. Some of the recommended verification products need to be produced using the CHPS display capabilities, whereas IVP and EVS can be used outside CHPS to produce a subset of the verification products. Therefore the schedule would be different for the CAT RFCs and the CAT-II RFCs. First the CAT RFCs would start developing the verification standards using the CHPS display capabilities, and share their progress with the CATII RFCs. The CATII RFCs would develop the standard verification products using CHPS when their CHPS implementation is being finalized (~late FY 2010).

October 2009: Finalize the second team charter.
October 2009 - March 2010: Perform user analysis to identify user requirements for CHPS-VS and test new prototype functionality (e.g., EVS version 2.0).
April - July 2010: Define requirements for the NWS Performance Branch and the RFCs to disseminate RFC verification information.
October 2009 - April 2011: Perform RFC verification case studies using IVP or EVS software as well as CHPS capabilities to test and improve the verification standards; continue to review/test new prototype functionality for CHPS-VS.
FY2011: Conduct third RFC verification workshop.
May - September 2011: Develop the final team report on improved verification standards and verification case studies; finalize and document the prototype capabilities developed at OHD and at the RFCs to produce the standard verification products.
September 2011: Final report due with verification case studies and prototype capabilities to produce standard verification products.