National Weather Service
River Forecast Verification Plan

Report of the
Hydrologic Verification System Requirements Team
October 2006

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

National Weather Service
River Forecast Verification Plan

Report of the
Hydrologic Verification System Requirements Team

October 2006

U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary

National Oceanic and Atmospheric Administration
Vice Admiral Conrad C. Lautenbacher, Jr., Administrator

National Weather Service
David L. Johnson, Assistant Administrator

Preface

The National Weather Service (NWS) River Forecast Verification Plan was developed by a team made up of representatives from NWS regions and headquarters. The team was chartered in September 2005 and commenced their work November 2005. The team was given the mission to establish the requirements for a comprehensive national river forecast verification. The team developed system requirements for river forecast center (RFC) operations, but the system can be applied to river forecasts in general. The team defined a system that is flexible and robust enough to meet NWS program management and scientific goals.

This report contains the team's recommendations for a river forecast verification plan as well as an implementation guideline that can be used for program formulation. The Plan developed by the team is based on the research efforts of Dr. Edwin Welles. His contributions to the team were critical in defining the system and setting key goals for a unified NWS river forecast verification program.

It is the belief of the team that our Plan will allow managers and scientists to make informed river forecast development decisions and resource allocations in the future.

The Plan is designed to be incorporated into the NWS requirements processes (OSIP and HOSIP) and provide a roadmap for future Advanced Hydrologic Predication Service projects and Water Resource Initiatives.

Peter Gabrielsen
Team Leader
October 2006

Table of Contents

Team Members

Peter Gabrielsen, Team Leader
Chief, Hydrologic Services Division, Eastern Region

Julie Demargne
Project Scientist, Office of Hydrologic Development, NWS Headquarters

Bill Lawrence
Development and Operations Hydrologist, Arkansas Basin River Forecast Center

Scott Lindsey
Senior Hydrologic Forecaster, Alaskan Pacific River Forecast Center

Mary Mullusky
Hydrologist, Office of Climate, Water, and Weather Services, NWS Headquarters

Donna Page
RFC Development Manager, Office of Hydrologic Development, NWS Headquarters

Noreen Schwein
Water Services Program Manager, Central Region Headquarters

Scott Staggs
Senior Hydrologist, California Nevada River Forecast Center

Tom Adams
Development and Operations Hydrologist, Ohio River Forecast Center

Kevin Werner
Hydrology Science Program Manager, Scientific Services Division Western Region

William Marosi
Hydrologist, Middle Atlantic River Forecast Center

Technical Advisor

Edwin Welles
Branch Chief, for the Systems Engineering Center Development Branch, Office of Science and Technology, NWS Headquarters

Executive Summary

In September 2005, the Hydrologic Verification System Requirements Team was formed to develop a River Forecast Verification Plan which defines a national river forecast verification system that can be used to make educated decisions to improve the NWS river forecast program. The team focused on the river forecast processes at the RFC, but the recommendations can be used to verify any river forecasts.

The team was charted to meet NWS related items in the NOAA Audit Action Plan. The team charter follows:

Vision: Provide easy access to enhanced river forecast verification data which will be used to improve our scientific and operational techniques and services.

Mission: Assess forecaster, program managers and user needs for verification data. Inventory current national and regional verification practices and identify unmet needs. Establish requirements for a comprehensive national system to verify hydrologic forecasts and guidance products which satisfy these needs. This system should identify sources of error and skill in the forecasts across the entire forecast process.

Scope of Authority/Limitations:

- Team will review the Gary Wick report and briefing entitled "Evaluation of Potential Forecast Accuracy Performance Measures for the Advanced Hydrologic Prediction Service."
- Team will review current HOSIP documents
  - Ensemble Verification and Validation NID-05-26 SON-05-001
  - Complete Deterministic Verification NID-05-016-SON-05-001
- Team should consider new verification science and methodologies for inclusion in the system
- Past or current practices, and organizational allegiances among the team members, must not be allowed to influence either the evaluation or the recommendations
- Team will consult with internal and external partners and customers as needed
- Team leader will have 51% of the vote and serve as team facilitator
- The team will make decisions by consensus if they can not meet by consensus the Team Leader can use 51% of the vote to make the decision.
- The team will solicit/incorporate minority opinions if decisions are not reached by consensus
- o travel expenses will be authorized

Termination Date: The team will be formed and commence activities by September 30, 2005 and complete their work NLT June 30, 2006.

Success Criteria/Deliverables: Deliver a NWS river forecast verification plan which measures skill and error in the forecast process. The plan includes conceptualized solution and a definition of operational requirements (through HOSIP Gate 2).

The River Forecast Verification Plan defines the roles and design of the verification system as well as the utility of verification. Verification has administrative and scientific uses and can be used to evaluate program performance, resource allocation and direct and improve scientific research. Verification is also useful to external partners and customers who can use it to understand the reliability and skill of river forecasts.

The plan defines seven categories of verification statistics that apply to both deterministic and probabilistic forecast verification, even though the metrics for the categories may be different. The categories include: categorical, error, correlation, distribution properties, skill scores, conditional statistics, and statistical significance. The plan includes recommended statistics for each of these categories that along with logistical metrics define a National Base Line Verification System (NBVS) that provides a sufficient framework for administrative verification. For scientific verification purposes, expanded statistics are defined and referenced for users who need to understand errors and compare current and newly developed forecast methodologies.

There are six components of verification system requirements. The verification system components include: selection of the forecast to be verified, identification of the skill and error sources, computation of the metrics and display of the verification results, archiving capability, hindcasting capability, and dissemination of the verification results and training. These requirement components apply to deterministic, ensemble and water supply forecasts.

The definition of the system incorporates as many existing tools that are available and provides a direction for incorporation in to the NWS requirements process. Besides leveraging the NWS requirements process there are a number of additional recommendations including soliciting peer review of our proposed verification system, identifying specific verification duties for RFCs and OHD, training, and defining uses for information provided by the verification categories and metrics.

Finally, the Plan identifies current verification activities that are taking place, identifies items that should be accomplished in the short term (FY07) and activities that will be completed in the future (FY08)– FY11) to support the Plan and how they fit into to NWS requirements process and the Advanced Hydrologic Prediction Services.

Introduction

River stage forecasts and more generally hydrologic forecasts produced by NOAA's National Weather Service (NWS) form the basis for decision making, including activation of emergency services for forecast flood events, reservoir operations for water supply, stream flow regulation, and recreational outings on the nation's streams and rivers. These hydrologic forecast, with forecast horizons that vary from several hours to several months, are produced in real-time using meteorological and hydrological data as input to sophisticated atmospheric, hydrologic, and hydraulic models. NWS hydrologists use these data and model output as guidance in producing their forecasts. Since these forecasts are used by the public to make critical decisions, it is essential that forecast performance is known and that forecast skill is measured to identify procedures in the forecast process that can be improved.

Forecast verification or measurement of the forecast skill has been challenging for the NWS largely due to the computational resources required. As a result, users of hydrologic forecasts have had few tools with which to assess confidence in those forecasts. Managers and other decision makers who allocate resources for hydrologic research and operational development have had little in the way of objective guidance upon which to base their decisions. For the management of the hydrologic program, the limited verification work done so far did not fully assess and validate existing and new forecasting components of the system, and did not help prioritize improvements necessary to realize reliable and skillful forecasts.

With the government oversight emphasis on objective performance management, the NWS is now more determined to verify hydrologic forecasts, taking into account every aspect of the forecasting process and source of skill and error to improve the hydrologic forecasts. Other recent scientific advancements include the addition of short-term and long-term ensemble forecasting for many rivers in the country. As with the traditional deterministic and statistical forecasts, these ensemble forecasts must be quantitatively verified to show their value and identify needed improvements.

To address this challenge, the NWS has committed to provide the means to verify hydrologic forecasts. The first step is to establish requirements for a comprehensive national system to verify hydrologic forecasts and guidance products which satisfy these needs. This includes identifying forecaster, program manager, scientific researcher and user needs for verification data. The system needs to verify the various forecast types currently produced by the NWS, including deterministic, statistical, and ensemble forecasts.

The Hydrologic Forecasting Process and its Error Sources

1. The River Forecast Center hydrologic forecasting process

In order to define the required verification system and identify sources of error across the entire forecasting process, that process first needs to be defined. Figure 1 shows the different steps of the forecasting process followed by the river forecast centers (RFCs) in producing hydrologic forecasts.

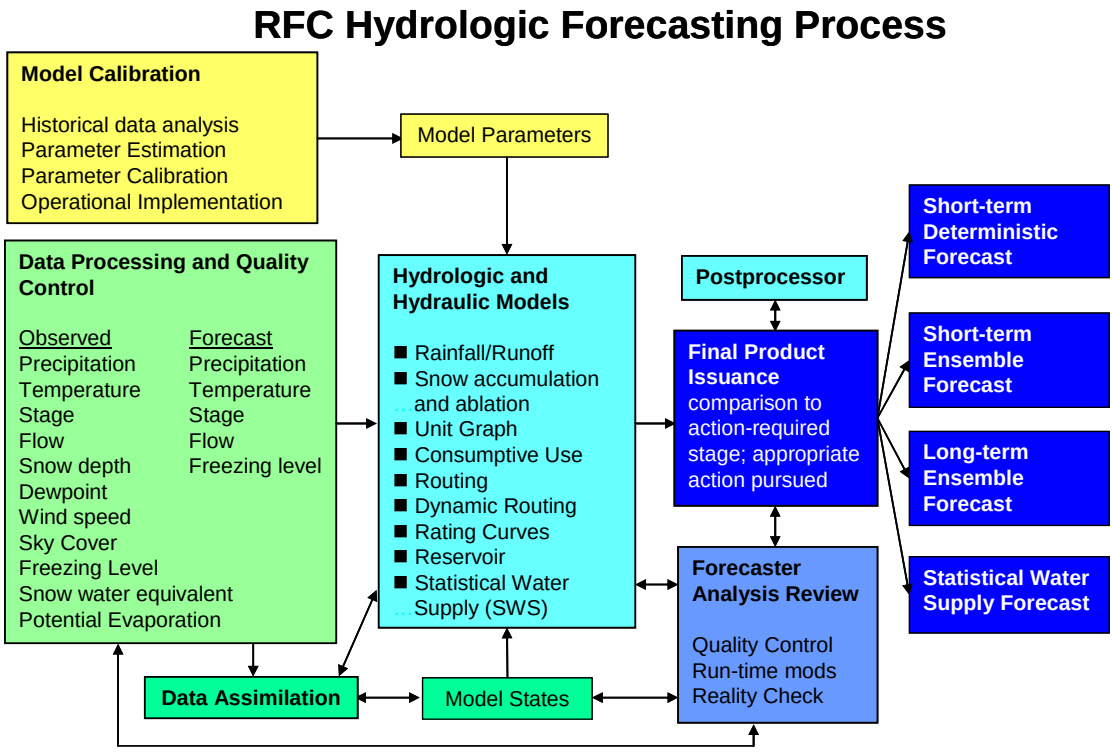

Figure 1 RFC hydrologic forecasting process

The detailed forecasting process shown in Figure 1 can be divided into four processing steps: (1) model setup (2) state updating, (3) forecast computation, and (4) product review and issuance.

Model setup is the process of selecting, parameterizing and linking a suite of models to simulate the hydrologic system to be forecast. Historical data are collected, quality controlled and corrected as needed. A set of simulation characteristics are selected (e.g., peaks, volumes, baseflow) and the models are tuned, by adjusting their parameters, to match the characteristics of the observed hydrograph record as well as possible. In some cases, the tuning is done manually, in other cases, it is done automatically using optimal search algorithms and objective functions.

State updating refers to the process by which the model states are adjusted to ensure the models have the best possible initial conditions with which to begin forecasting. To conduct the data assimilation, observations are collected, quality controlled and fed to the

models during the observed period; the model parameters, the model states, and/or the input time series are then adjusted to make the model performance match the observations. The actual adjustments and search for the best fit to the observations may be done manually (as currently done at the NWS) or automatically through statistical techniques.

Forecast computation is when observed and forecast data are used to drive the hydrologic and hydraulic models into the future to some desired time. It could also include a post-processor component to account for the hydrologic uncertainty and improve the forecast. A variety of input forecasts may be included depending upon the methods employed to model a basin's hydrologic system. In most cases, Quantitative Precipitation Forecasts (QPF) will be used; although temperatures, reservoir releases, stream flow regulation, lock and dam schedules or upstream flows can also be important. In some cases, the skill of the local hydrologic modeling will be overwhelmed by the skill (or non-skill) of the input forecasts, and verifying the hydrologic forecasts becomes a matter of verifying the input forecasts.

Product review and issuance is done by the forecaster. The hydrologic forecaster reviews the model output and constructs a final forecast. The human quality control of the forecast is critical to the forecasting process as computational procedures may arrive at unrealistic solutions to simulations of complex hydrologic systems.

2. Error sources in the hydrologic forecasting process

Errors are introduced throughout the whole hydrologic forecasting process. While some of these errors, such as gage maintenance and snow depth, typically have relatively minor effects on the final hydrologic forecast product, other error sources, such as calibration, snow water equivalent and QPF, can have more significant ramifications. Some of these error sources may only affect the forecasting process seasonally or occasionally. Additionally, while many error sources may be considered relatively "minor," an accumulation of these "minor" errors may ultimately prove to be significant. Also research has shown that the hydrologic errors and the hydrometeorological errors are not simply additive but interact with each other (Welles 2005 and reference herein Krzysztofowicz 1999).

The error sources are mainly:

- the input data, which includes errors from observed data, forecast data and climate outlooks, rating curves, and stream flow regulation outflows and releases (including lock and dam);
- the hydrologic and hydraulic models, which include errors from model parameters, model states, and model structure;
- the forecaster analysis.

Details about these error sources are given in Appendix 1. Different processes, both automated and manual, are developed operationally and experimentally to reduce the impact of these errors, as shown in Figure2.

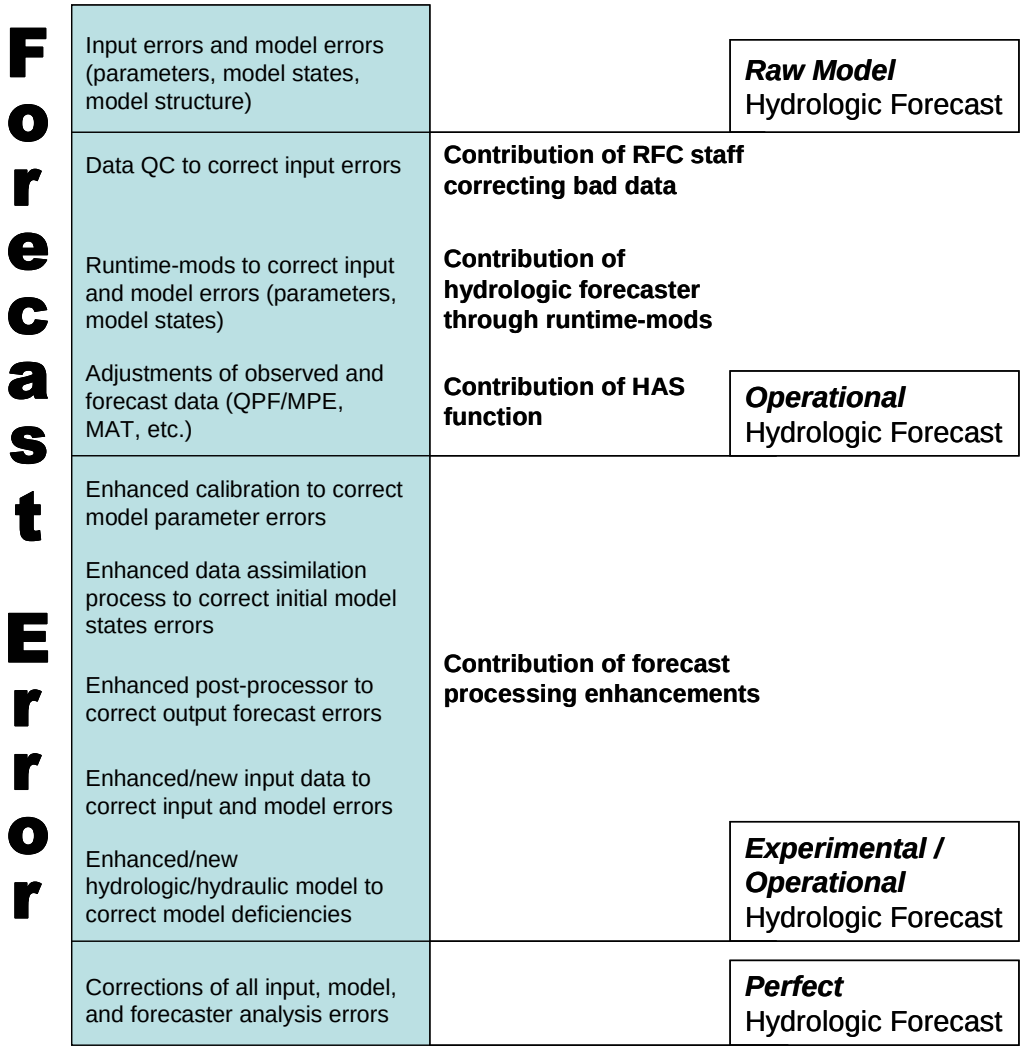

Figure 2 Impact of various forecasting processes on the forecast error, from raw model forecast up to perfect forecast

The raw model hydrologic forecast corresponds to the forecast generated by running only those parts of the forecasting process that are automated, requiring no interaction with the forecaster during operational use. The raw model forecast has errors coming from the input data and the forecast models.

During the operational forecasting process, the forecaster aims at reducing these errors with three different human processes:

- the data quality control to eliminate some of the incorrect input data;
- the runtime modifications to modify the input data as well as the model parameters and model states;
- the adjustments of observed and forecast data.

While reducing the input error and model error, these processes also introduce an additional error from the forecaster analysis. The resulting forecast is the operational

forecast whose errors come from the input data, the forecast models, and the forecaster analysis.

To further reduce the errors from input data, forecast models, and forecaster analysis, several processes are or could be developed and enhanced. Any of these processes could be integrated in the operational forecasting process if it improves the forecast performance; so the experimental forecast generated with this additional/enhanced process would become the new operational forecast. These forecasting processes are:

- a re-calibration of the hydrologic and hydraulic models to reduce the model parameters error;
- a data assimilation process to get better estimates of the initial model states, thus reducing the model states error;
- a post-processor component to reduce the hydrologic error of the output forecast;
- the integration of new or enhanced input data (from new sources for example) to correct some of the input errors and model errors;
- the integration of new or enhanced hydrologic or hydraulic model to reduce the model error, especially the structural error.

All these forecasts need to be evaluated with a perfect forecast (with no input error, model error, and forecaster analysis error), which is known retrospectively through observations.

Verification Purposes

1. Role and setup of the verification system

The verification system aims at monitoring the forecast quality over time. Verification helps improve the forecast quality by knowing the strengths and weaknesses of the existing forecasting system and by comparing the quality of different forecasting methodologies. The forecast performance should be evaluated at various steps during the forecasting process, as shown below in Figure 3. Although, each step in the forecast development process is assumed to contribute skill to the final forecast product, the individual contributions of the forecaster, the forecast models and the input data to the overall skill are not known, nor can they be known without comprehensive verification.

In order to analyze the different error sources, the verification system needs to evaluate the forecasts generated by the entire forecasting process as well as the forecasts generated by the individual forecasting processes at intermediate steps. The comparison of the forecast performance with and without a specific process will help understand the relative impact of that process on the forecast quality. For example, to assess the contribution of the data assimilation process, the verification system needs to evaluate the forecasts without and with the data assimilation process. Running the verification system at the intermediate steps of the hydrologic forecasting process will determine whether the individual forecasting processes lead to a performance gain. Also it will help prioritize the enhancements of the individual processes necessary to generate reliable and skillful forecasts.

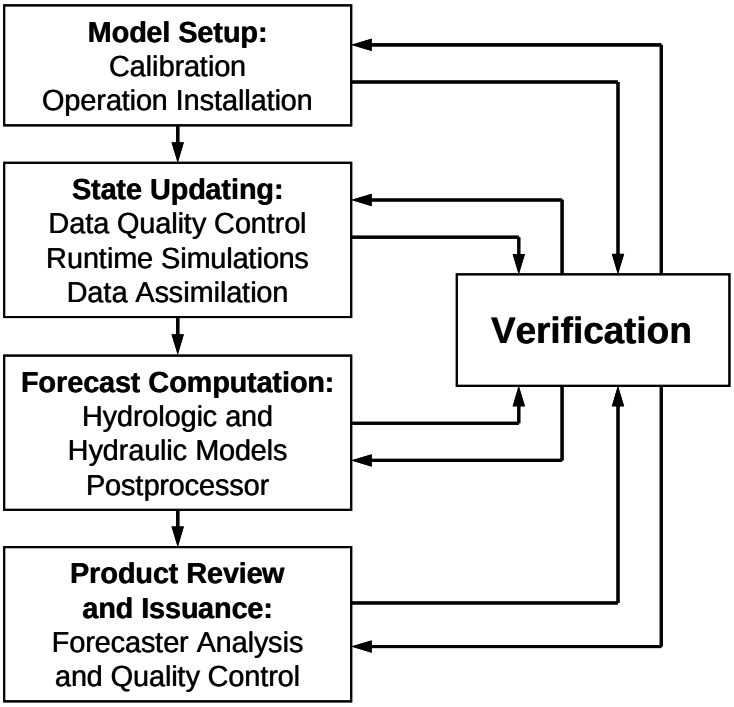

Figure 3 Setup of the verification system to evaluate the processing steps of the forecasting system

2. Multiple uses of the verification results

An assessment of how customers might use forecast verification information is a vital step in the process of designing tools to measure forecast accuracy and skill. Customers may include hydrologic program managers, emergency managers, scientists and researchers, hydrologic forecasters, and everyday users of hydrologic information. These customers may use the verification results in two different modes: the operational mode and the experimental/research mode. The operational mode aims at determining whether or not operational forecasts are reliable and skillful and if they enable better decision making relative to alternatives, such as climatology, persistence, etc. The research mode aims at identifying the sources of skill and error in the forecast system and comparing the performance of various forecasting processes to determine what processes lead to the best forecast performance. Some elements of skill and error may be analyzed through the operational framework, but typically at a cost of the operational time needed to create scenario operational forecast runs, large archive storage needs, and the years of data needed before meaningful analysis can be performed. To meet the multiple user needs, the verification system should describe the performance of: past and recent forecasts; operational forecasts, control (or baseline) forecasts, and experimental forecasts; forecasts for a certain time period or relative to specific events. A description of the multiple uses of the verification results follow.

Hydrologic program managers might use forecast verification to show program value and program improvement as well as to make important decisions on scientific and resource investment. In a potential flood situation, an emergency manager might use forecast verification metrics to assess confidence and skill in a flood forecast to make decisions on early deployment of emergency personnel and equipment. Hydrologic researchers can look at forecast verification to determine sources of error in hydrologic forecasts that in turn indicate research areas that have the most value to the science of hydrologic modeling. RFC forecasters might use forecast verification to identify personal biases in the forecasting process. An example might be consistently under-estimating the peak stages when forecasting a rain-on-snow flood event. Verification metrics will also help forecasters target the portions of the forecasting process that will produce the greatest benefit in forecast accuracy and skill. If the greatest error source in a stage forecast is data quality, forecasters may spend more time on data quality control and assurance and less time making run-time modifications to the forecast. If the greatest error source turns out to be model calibration, the research community could also be engaged to develop better calibration methodologies.

Finally, through hydrologic forecast verification, everyday users of hydrologic forecasts can begin making informed decisions based on indicated forecast skill and accuracy. For instance, a rafting guide operator may choose to close for a day based on a forecast of dangerously high stages that has proven through experience and published verification metrics to be accurate. In another scenario, fishermen may decide to hit the water early based on a forecast water level that is favorable to their success and experience that the forecast is accurate. Finally, a dam operator can make informed decisions using concepts of risk analysis and forecast skill to reduce reservoir releases during the snowmelt season based on long-term probabilistic forecasts of snowmelt runoff in areas where water supply is an important issue. Some of these customers might look at verification statistics for a number of months or years to determine their confidence in a forecast. Others may look at the statistics over the most recent days or weeks to make informed decisions. But in all cases, access to robust forecast verification information can help users make informed decisions that will better fulfill the NWS mission to protect life and property and to enhance the national economy.

Characteristics of the Verification System

1. Administrative and scientific verification

Because of the variety of the verification purposes, Brier and Allen (1951) have differentiated between *administrative* verification and *scientific* verification. *Administrative* verification is descriptive, providing characterizations of the status of the forecast service, such as the timeliness of the forecast delivery, the number of forecasts issued, and, the overall forecast skill. The goal of the *administrative* verification is to describe the efficiency of the forecast service and the overall forecast performance so decisions can be made with respect to resource allocation, research directions and implementations strategies (Welles 2005). The administrative component of the verification system provides logistical hydrologic verification measures to describe the quality of the forecast service as well as a few forecast skill measures to describe the

overall forecast performance. The *scientific* component of the verification system aims at analyzing all the aspects of the forecast skill and reliability. Its purpose is to identify the sources of skill and error in the forecasting system and their impact on the forecast quality (see Figure 3) so enhanced forecasting processes could be developed to improve the forecast performance. The requirements for a comprehensive verification system presented in this report are relevant to both *administrative* and *scientific* verification. This comprehensive verification system includes a National Baseline Verification System (NBVS) to be defined on a national scale for *administrative* verification purposes.

2. Logistical and forecast skill verification measures

To evaluate the quality of the forecast service as part of the *administrative* verification, the hydrologic verification system needs to provide logistical measures of the forecast service that measure non-skill attributes of the delivered forecasts. The purpose for collecting the logistical information is to answer questions like the following. What new types of forecasts have been developed? Is the number of forecast locations increasing or decreasing? Have computational improvements reduced the effort to issue a forecast? Have methodological improvements reduced the time it takes to prepare a basin for forecasting?

Besides these logistical verification measures, the hydrologic verification system needs to provide verification metrics to assess the quality of the forecast as part of both *administrative* and *scientific* verification. Characteristics of the forecasts that are evaluated with verification metrics are:
Reliability - when an event is forecasted, are the forecasts reliable?
Discrimination or resolution - do the forecasts distinguish between the types of upcoming events?
Accuracy - the level of agreement between the forecast and the truth represented by observations.
Error - the difference between the forecast and the observation.
Skill - the relative accuracy of the forecast over some reference forecast generally an unskilled forecast such as persistence or climatology due to the forecast system itself.
Association - the strength of the linear relationship between the forecast and the observation.
Bias - the correspondence between the mean forecast and mean observation.
Uncertainty: - the variability of the observed variable.

Categories of verification metrics were developed by the team to ensure the verification system is broad enough to capture the different aspects of verification and meet the multiple needs of the users. The seven categories established from Welles (2005) apply to both deterministic and probabilistic forecast verification, while the metrics themselves may be different.

Categorical: statistics related to predefined threshold or range of values (e.g., above flood stage, minor).

Error: statistics that measure various differences between forecast and observed values (including timing errors).

Correlation: statistics that measure the correspondence between ordered pairs (e.g., crest forecasts vs. QPF, forecast and observed stages).

Distribution Properties: statistics that summarize the characteristics of a set of values.

Skill Scores: statistics that measure the relative accuracy with respect to some set of standard reference or control set of forecasts.

Conditional Statistics: metrics computed based on the occurrence of a particular event or events such as a specific range of observations or forecasts.

Statistical Significance: measures the uncertainty of the computed values of verification metrics.

Regarding the selection of verification metrics, the team recognized that it is especially crucial to define a few program-level performance measures for program managers to track over time the overall forecast performance and to demonstrate improved efficiencies or cost effectiveness of the forecast service on a national scale. In Welles et al. (2002), the authors underlined the difficulty to determine which verification statistics were best suited to be aggregated on a large number of forecast points to characterize the forecast quality on a national basis. Also Wick (2003) evaluated potential forecast accuracy measures for the Advanced Hydrologic Prediction Service (AHPS) for both deterministic and probabilistic forecasts. In this report, Wick emphasized the limited amount of data currently available to compute the verification statistics and the variety of forecast performance characteristics relevant to the hydrologic program. Therefore he underlined the need for further evaluation of multiple verification metrics with more data to select the best program-level performance measures. The development of the National Baseline Verification System, including the adequate capabilities of archiving, hindcasting, and objective assessment of the forecast performance as described in the system requirements, will meet that need. Experience in evaluating forecast performance with more data through a large number of verification metrics will lead to define which measures for program-level performance evaluation as well as other verification purposes should be used.

3. Verification metrics

Verification metrics consist of logistical measures to assess the quality of the forecast service and forecast skill statistics to assess the quality of the forecast.

For a comprehensive description of the service efficiency, the following logistical measures are required:

- characterizing point forecasts by service type, frequency and location;
- characterizing areal forecasts by service type, frequency and location;
- identifying daily the number of issued forecasts by type and location;
- quantifying the person effort required to set up a basin for forecasting, including data gathering, calibration, model setup and implementation efforts;

- quantifying the person effort required to issue each type of forecast, including manual quality control of input data, forecaster run-time modifications and forecaster review and analysis;
- quantifying the timeliness of issued forecasts.

The end goal requirement is to standardize and automate the collection of these logistical verification measures and provide a national database of those measures to those who manage the hydrology program. Query tools should be provided, such that managers may query the national database of logistical measures and create meaningful assessments.

Regarding forecast skill verification measures, most commonly used verification metrics for the seven categories are provided in Table 1. They describe the different aspects of forecast verification. They should be integrated in the verification system to meet the multiple needs of the end users. The selection of metrics depends on the verification purposes of the users. At this point, it seems necessary to incorporate multiple verification metrics in the baseline verification system to compute these metrics on large datasets and therefore better understand the forecast performance; then it would be possible to determine which metrics should be used for which verification purposes. For *administrative* verification purposes, the overall quality of the forecast could be described with a few metrics highlighted in bold red in Table 1, defined in Appendix 2, and provided with detailed definitions and examples in Appendix 3. These metrics along with the logistical measures make up the National Baseline Verification System. For *scientific* verification purposes, the combined set of metrics would be available to the forecasters and the other users, such as the scientists who need to understand errors and compare current and newly developed forecast methodologies. The definition of these verification metrics are provided in several verification books and websites such as Wilks (1995), Franz and Sorooshian (2002), Joliffe and Stephenson (2003), and WMO (2004). Since research in the verification domain aims at developing new verification metrics (especially user-oriented verification measures), the system should allow for future inclusion of more sophisticated measures to better characterize the forecast performance for specific end users or verification purposes.

CATEGORIES	DETERMINISTIC FORECAST VERIFICATION METRICS	PROBABILISTIC FORECAST VERIFICATION METRICS
1. Categorical	**Probability Of Detection (POD), False Alarm Rate (FAR)**, Critical Success Index (CSI), **Lead Time of Detection (LTD)**, Pierce Skill Score (PSS), Gerrity Score (GS)	**Brier Score (BS), Rank Probability Score (RPS)**
2. Error (Accuracy)	**Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Error (ME), Bias (%)**, Linear Error in Probability Space (LEPS)	Continuous RPS
3. Correlation	**Pearson Correlation Coefficient**, Ranked correlation coefficient, scatter plots	
4. Distribution Properties	Mean, variance, higher moments	Wilcoxon rank sum test, variance of forecasts, variance of observations, ensemble spread, Talagrand Diagram (or Rank Histogram)
5. Skill Score	**Root Mean Squared Error Skill Score (SS-RMSE) (with reference to persistence, climatology,** lagged persistence), Wilson Score (WS), Linear Error in Probability Space Skill Score (SS-LEPS)	**Rank Probability Skill Score**, **Brier Skill Score (with reference to persistence, climatology,** lagged persistence)
6. Conditional Statistics	**Relative Operating Characteristic (ROC) and ROC Area, reliability measures, discrimination diagram**, other discrimination measures	**ROC and ROC Area**, other resolution measures**, Reliability diagram, discrimination diagram**, other discrimination measures
7. Confidence	**Sample size**, Confidence Interval (CI)	Ensemble size, **sample size,** Confidence Interval (CI)

Table 1 Verification metric categories including common metrics for deterministic and probabilistic forecasts. The metrics highlighted in bold red are recommended to be part of the National Baseline Verification System.

In addition to analyzing the deterministic forecasts without uncertainty, they should be be converted to probabilistic form by overlaying an estimated error distribution around the deterministic forecast value. An easy and reasonable error distribution comes from the calibration statistics. The seasonal water supply outlooks have already been issued in this form for years (with their 90,70,30, and 10% exceedance quantiles). This type of

conversion will provide additional forecast information and allow direct comparison with deterministic forecasts.

4. Review of available verification tools

Currently the NWS RFC hydrologic forecast verification program is disjointed. However, there are a number of regional verification techniques available and one national verification scheme. At this point the national verification scheme for RFC hydrologic forecasts only computes error statistics (root mean square error, mean absolute error and mean algebraic error) for deterministic river forecasts. The following review of verification tools aims at developing from the current experiences with existing tools an enhanced and comprehensive verification system.

The 13 RFCs were surveyed to determine the type of verification currently being done at their offices and the tools used to do so. The team also reviewed other verification tools developed by related organizations and contractors, such as OHD and universities. Table 2 lists some of the tools available or being developed, the output metrics of each, the forecast type the tool currently processes (including deterministic, ensemble and statistical water supply), and the organization responsible for development and maintenance of the tool. It is recognized that additional external research and partnered research may develop additional verification tools.

Verification Tool	Output Metrics	Forecast type processed	Responsibility
SR Categorical system	POD, FAR, LTD; weighted for normalization of RFCs across region	Deterministic	SR
Interactive Verification Program (IVP)	POD, FAR (traditional and hydrologic), Under Forecast Rate, Over Forecast Rate, CSI, LTD, RMSE, MAE, ME, Maximum Error, Scatter Plots, Quantiles and Extremes	Deterministic	OHD
University of Iowa AHPS probabilistic verification	Bias, Skill, Potential Skill, Conditional Bias, Unconditional Bias, Skill Score	Long-term Ensemble	University of Iowa
Ensemble Verification Program	BS and its decomposition, RPS, BSS & RPSS (vs. climatology and persistence), ROC, Reliability Diagram, Scatter Plots, and for ensemble mean, Bias Ratio, ME, RMSE, Correlation Coefficient	Short- to medium-term Ensemble (lead days 1 to 14) with climatology and persistence as reference forecasts	OHD
R Verification	BS, LEPS, Contingency Tables, ROC, Reliability Plots, Continuous RPS	Short- and Long-term Ensemble, Deterministic, and Statistical Water Supply	freeware
NERFC Verification	POD, FAR, CSI	Deterministic	NERFC
ABRFC Verification	RMSE, Bias	Deterministic	ABRFC
CNRFC Water Supply (WRH)	Percent Error, Relative Error	Statistical Water Supply	CNRFC
CNRFC Forecast Stage Verification	Skill (vs. persistence), Error, Bias, Maximum Deviation Detection for Forecast Points and Groups	Deterministic	CNRFC

Table 2 Verification tools available or under development

Currently there are four verification projects in various stages of the NWS Requirements Process (HOSIP and OSIP). These existing projects will be used to define all the RFC hydrologic verification system requirements that are needed for the comprehensive verification system; Table 3 summarizes the projects and status.

Project	Identification	Status
Logistical Hydrologic Verification Measures	HOSIP: NID 06-008 – SON-06-001	G1: Approved G2: Conditional Approval; incorporated comments from gate meeting; electronic Gate 2 approval is pending G3: Scheduled for September 27, 2006
	OSIP: 06-030	G1: Approved G2: Approved – redirected to SREC
Hydrologic Deterministic Verification	HOSIP: NID-05-016 – SON-05-001	G1: Approved G2: Approved G3: Approved G4: Scheduled for 03/08/06 (delayed, will schedule for June 2006)
	OSIP: 06-023	G1: Approved G2: Approved – redirected to SREC
Hydrologic Ensemble Hindcaster	HOSIP: NID-05-024-SON-05-001	G1: Approved G2: Approved G3: Need to Schedule ~ September 2006
	OSIP: 06-024	G1: Approved (put on SREC list) G2: Need to schedule ~ September 2006
Hydrologic Ensemble Verification & Validation	NID-05-026-SON-05-001	G1: Approved G2: Approved G3: Need to Schedule ~ September 2006
	OSIP: 06-025	G1: Approved (put on SREC list) G2: Need to schedule ~ September 2006

Table 3 Status of projects in the NWS requirements process

Verification System Requirements

The team has identified six components of verification system requirements. The verification system components include: selection of the forecast to be verified, identification of the skill and error sources, computation of the metrics and display of the verification results, archiving capability, hindcasting capability, and dissemination of the verification results and training. These requirement components apply to deterministic, ensemble and water supply forecasts and must be weaved into the appropriate NWS Requirements documents to be implemented in different phases of development.

1. Selection of forecast to be verified

The verification system shall provide the ability to verify forecast relative to the different variables used in the hydrologic forecasting system: forcing input variables (mainly precipitation and temperature) and hydrologic variables (flow and stage). The user should then define the time interval and statistical variable of the forecast to be verified if necessary. For example, verification could be done on 6-hr or 24-hr flow forecasts, as

well as weekly or monthly flow volume forecasts; it could be relative to the mean, median, maximum or minimum value during that time interval.

The system should allow the user to select the set of forecasts for which the computed verification metrics would be statistically significant and informative. For determining the forecast to be verified, the user needs to find a trade-off between two constraints. The sample needs to be large enough to compute robust verification metrics. At the same time, the sample needs to be small enough to be homogeneous and distinguish the different factors that contribute to the forecast error. This process of forecast stratification and selection depends upon the purposes of verification; so the verification system should account for the needs of all the end users. The verification system should be flexible to allow forecast stratification and selection according to:

- time attributes (days, months, seasons, years, as well as lead time)
- service attributes (national, regional, RFCs, groups, locations)
- individual forecaster within guidelines agreed to by the NWS and the NWSEO
- basin attributes (response time, size, slope, aspect, elevation, snow, non-snow)
- forecast or observed events (crest timing, rising and falling hydrographs)

2. Analysis of skill and error sources

The verification system should provide the ability to identify the sources and sinks of forecast skill through the inter-comparison of the performance of multiple forecast scenarios. The sets of forecasts from these scenarios should be archived and then processed by the verification system. The verification results from these forecast sets should then be inter-compared to pinpoint sources of error and skill in the forecasting process. This work is crucial to assess the contribution of existing and enhanced forecasting processes on the forecast error as described in Figure 2. Many of the comparison studies will be undertaken by research scientists rather than operational or developmental entities to determine what processing steps and methodologies are best for the operational forecasting system. However the conclusions of this diagnosis work should be available to all end users including forecasters and developers.

The analysis of skill and error sources should be done for the three error sources: input meteorological and hydrologic data, hydrologic and hydraulic models, and input from human forecasters. For the input data error, the quality of the input data should also be evaluated. Below are examples of scenarios to be run to generate the various hydrologic forecast sets required to analyze each of these error sources.

Scenarios for analyzing the impact of input data errors on the hydrologic forecast:

- With QPF vs. other reference precipitation forecast (persistence, zero QPF), to determine the value of QPF relative to other reference forecasts
- With traditional point based MAP vs. MAPX from MPE, to determine the impact on the method used to compute precipitation
- With QPF vs. Perfect QPF (from observed data), to determine the impact of the QPF errors
- With FMAT vs. other reference temperature forecast (persistence), to determine the value of FMAT relative to other reference forecasts

- With FMAT vs. Perfect FMAT (from observed data), to determine the impact of the FMAT errors

Similar scenarios could be developed to test the impact of different forecasts for potential evaporation, freezing level, reservoir outflows and releases, as well as the impact of different observed data (from various data sources or methodologies).

Scenarios for analyzing the impact of model errors on the hydrologic forecast:

- With pre-recalibration parameters vs. post-recalibration parameters, to determine the value of the recalibration process
- With lumped parameter model vs. distributed model, to determine the value of distributed modeling
- With automated state updating vs. no state updating, to determine the value of state updating
- With post-processing vs. no processing, to determine the value of post-processing

Similar scenarios could be developed to compare the value of different methodologies for calibration, data assimilation, or post-processing.

Scenarios for analyzing the impact of forecaster analysis errors on the hydrologic forecast:

- With data QC vs. no data QC
- With run-time modifications vs. no run-time modifications
- With operational model (includes forecaster intervention) vs. raw model (without forecaster intervention)

The definition of the raw model needs to be determined. For the purpose of this plan, a raw model is defined to include only those parts of the forecasting process that are automated, requiring no interaction with the forecaster during operational use.

3. Computation of verification metrics and results presentation

The verification system should provide the ability to compute the verification metrics defined in Table 1, and at a minimum the subset of metrics recommended for the NBVS. For skill scores, the forecast performance is compared with the performance of another reference forecast. This reference forecast should include persistence, lagged persistence (trend), and climatology. It is also essential to compute all the verification metrics for both actual forecasts and control (or baseline) forecasts to compare the performance of actual forecasts with these alternative forecasts. It helps the user understand the magnitude of the computed metrics and provide a perspective on forecast performance (Welles 2005). Control forecasts include persistence, climatology, raw model forecast, and simulated hydrologic forecast from perfect forcing input; this simulated flow/stage forecast is particularly important to separate the errors from forcing input and the hydrological errors.

The verification system needs to be modular and flexible to integrate newly developed verification metrics or new reference forecasts, or other verification methods preferably using common statistical analyses packages such as R (http://www.r-project.org/index.html).

The verification system should also include a graphical capability to display the verification results for all the different metrics where it is appropriate. The display capability should be flexible to allow the user to analyze the verification results according to lead time, verification time window, spatial location, as well as type of variable to be verified (forcing input and hydrologic forecasts). It should allow both run-time displays as well as inter-comparison display to support the diagnosis work described in Section 2. To the greatest extent possible, displayed plots should be customizable by the user (e.g., plot titles, axis labels, legends). The data formats should also be flexible and agile enough to support new plotting functionality. Also, a capability to access all data included in any plot should be readily available from the plotting procedures.

4. Archiving

The need for standardized archive datasets in a common format is critical to the success of any verification system since the verification metrics need to be computed on a large sample of forecasts to be statistically significant. It is also important that specific datasets needed by particular verification metrics be saved in the same format by all RFCs. However RFCs have long lacked a standardized method of archiving products and graphics. Since available data for verification has been very limited so far, the forecast verification analysis at the NWS was restricted as Franz and Sorooshian (2002), Wick (2003), and Welles (2005) emphasized. With the arrival of the RFC Archive machines (RAX) in 2004, the effort to archive and standardize what is actually saved by each and every RFC has just started.

A standard archive database is an integral part of the verification system and should be defined, including data types and data formats, even before the verification system is operational. The archiving capability is especially crucial to store all the information generated with the intervention of forecasters since the processes involved are not automated. This includes archiving all the observed and forecast data used by the hydrologic forecasting system. The definition of the standard archive database should occur as soon as possible to get sample sizes as large as possible for forecast verification.

The archive tools for the verification system should store all the information necessary to stratify the forecast datasets in the different modes described in Section 1. Specifically, the archive database should include information that allows sorting by:

- Time attributes (days, months, years, seasons)
- Service attributes (national, regional, RFCs, forecaster, groups, locations)
- Basin attributes (response time, size, slope, aspect, elevation, snow, non-snow)
- Rising and falling hydrographs (observed and forecast)

The system should also be able to archive raw model data. Also it would be necessary to determine how to capture any new or enhanced forecasting process (see Figure 2), including any OFS runs that modify model parameters, segment definitions, and station or area definitions.

5. Hindcasting

The system should include the capability to hindcast/re-forecast all the forecast data and time series required for the diagnosis work described in Section 2, for which the forecasting scenarios do not include the forecaster intervention. It should produce hydrologic hindcasts, as well as forcing input hindcasts and retrospective model states over a time period up to multiple years. It should also provide real-time access to the available hindcast archive and robust metadata that fully describes each hindcasting scenario.

This hindcasting capability is crucial since verifying the hydrologic forecasts requires a sample size large enough for robust verification statistics. The hindcasts to be generated for a given forecasting scenario would reflect a single forecasting system, with no changes relative to the hydrologic and hydraulic models. Therefore a capability for routine, systematic and rigorous hindcasting is necessary to assess and validate any new forecasting process, including the enhanced processes described in Figure 2.

Similarly to the various scenarios described in Section 2, the hindcasting capability should generate hydrologic forecasts from:

- Different QPFs (e.g., Perfect QPF, zero, actual, persistence)
- Different FMATs (e.g., Perfect FMAT, actual, persistence)
- Different freezing levels
- Different MAPEs
- Different reservoirs forecasts
- Different stream flow regulation scenarios
- Different QPEs (e.g., point based MAP, MAPX, Q2)
- Different sets of model parameters
- Different models, including the post-processing and state updating models

Because of the numerous scenarios and the large potential quantity of data involved in this hindcasting work, modification of the functionalities used to store and archive all the data may be necessary.

6. Dissemination and Training

The system should allow disseminating the verification results along with the hydrologic forecasts to all the end users, as well as providing real-time access to data.

The system should include comprehensive documentation about the meaning of the verification metrics and the methods used to develop and analyze the verification results.

Recommendations

The team recommends the four OSIP projects defined in Table 3 be used as the vehicles to implement the verification system. The categories, metrics, baseline and expanded (administrative and scientific) systems are defined in this report.

Besides including the requirements in the OSIP projects mentioned in Table 3, there are a number of additional steps the team recommends to successfully implement the verification program. These include soliciting peer review of our proposed verification system, training, and defining uses for information provided by the verification categories and metrics. Listed below are a number of specific recommendations which can help develop an effective RFC river forecast verification program.

- OHD should assign a program manager for verification.
- Establish formal verification focal points at each RFC.
- Create a national team with the responsibility for defining, developing, and implementing standardized formats and procedures for archive data to support the verification system.
- Create national river forecast performance goals. This should be accomplished once the software has been fielded and some experience gained with the metrics.
- Ensure adequate hydrologic verification training, and use of the system, is captured in OSIP documentation.
- Publish findings of this report in peer reviewed journals (e.g., BAMS, EOS) to inform the research community of our plans.
- Ensure an end-to-end assessment and verification of the elements in the hydrologic forecasting process that are outside of the control of the RFC forecaster or produced by other agencies. Each element should be analyzed and verified (input data such as QPF, gridded elements QPE, rating curves, as well as output forecasts such as flash flood guidance).
- OHD needs to establish a team to define the raw model to enable the users to assess the impact of various steps (e.g., calibration, quality control, run-time modifications) on the forecast performance.
- Archive of necessary data to support verification software should begin within 30 days of the data being defined.
- Ensure continuity with other activities that support this verification plan.
- Brief the National Performance Management Committee (NPMC) and ensure incorporation of the RFC hydrologic verification requirements.

Development and Implementation

Verification System Implementation

The National Baseline Verification System aims at quantifying the quality of the RFC forecasts and the quality of the forecast service. Its goal is also to identify sources of error and skill in the forecasts across the entire forecast process. It includes logistical, deterministic and probabilistic components.

In order to achieve the deliverables stated above, a verification system must provide specific capabilities. Those capabilities are:

1. **Data archiving:** All data needed for a full verification system must be archived regularly.
2. **Computing metrics:** A tool must be made available to calculate the desired metrics.
3. **Displaying metrics:** Forecasters, scientists, and users must be able to examine the metrics, and this is best done through graphics and formatted reports.
4. **Disseminating the metrics and data:** The metrics must be disseminated to the public so that the end users can understand the quality and usefulness of the forecasts and for collaborative verification analysis.
5. **Real-time access to metrics:** The metrics must be made available in real-time, updated regularly, in order to allow forecasters to understand the errors both in recent forecasts (the past week) and over the long term (the past decade).
6. **Error analysis:** Forecasters must be able to fully analyze their forecasts in order to identify the sources of errors and compensate for them. This requires using multiple forecast scenarios, including hindcasting experiments, and analyzing the input to the forecast system, as well as the output.
7. **Performance measure tracking:** Performance measures must be produced, reported, and tracked, showing the level of success of RFC forecasting.

To facilitate better development of the NWS river verification program components (logistical, deterministic, probabilistic, and gridded) a specific list of milestones achieved during FY06 (items completed in FY06 are highlighted) and a detailed list of the proposed FY07 activities along with a general schedule of the verification activities through FY11 is provided.

To asses the quality of the forecast service, logistical verification measures should be defined in the verification system to measure non-skill attributes of the delivered forecasts. Using the seven capabilities described above, the status of the verification work for logistical verification is given below.

Logistical Verification Task Summary

Task	FY06	FY07	FY08	FY09	FY10	FY11
1. Archive						
2. Compute Metrics						
3. Display Metrics						
4. Dissemination						
5. Real Time Access						
6. Error Analysis						
7. Performance Error Tracking						

FY06

1. **Data archiving:** Not yet available. Work has just started to archive the information for characterizing point forecasts by service type and frequency.
2. **Computing metrics:** Not yet available. The envisioned metrics are: 1) characterizing point forecasts and areal forecasts by service type, frequency and location; 2) identifying daily the number of issued forecasts by type and location; 3) quantifying the timeliness of issued forecasts; 4) quantifying the person effort required to set up a basin for forecasting, including data gathering, calibration, model setup and implementation efforts; 5) quantifying the person effort required to issue each type of forecast, including manual quality control of input data, forecaster run-time modifications and forecaster review and analysis.
3. **Displaying metrics:** Not yet available.
4. **Disseminating the metrics and data:** Not yet available.
5. **Real-time access to metrics:** Not yet available.
6. **Error analysis:** Not yet available.
7. **Performance measure tracking:** Not yet available.

FY07

- Propose a plan to archive the required information for logistical measures (Item 1).
- Propose a plan to compute/get all the logistical measures. Start implementing the logistical measures (Item 2)

Forecast verification

There is a need to verify three types of hydrologic forecasts generated by RFC forecasters: deterministic, probabilistic, and grid forecasts. Using the seven capabilities described above, the status of the verification work for each of the three kinds of forecasts is given below.

Deterministic Verification Task Summary

Task	FY06	FY07	FY08	FY09	FY10	FY11
1. Archive						
2. Compute Metrics						
3. Display Metrics						
4. Dissemination						
5. Real Time Access						
6. Error Analysis						
7. Performance Error Tracking						

FY06

1. **Data archiving:** Currently available at RFCs via the archive database on the RAX machines but it must be evaluated to determine whether all the data needed for verification and hindcasting (including forecast attributes to stratify forecast when computing verification statistics) are actually stored.
2. **Computing metrics:** Some metrics can be calculated via the current verification software (IVP) for stage data; need to make computations available for more data types (flow, precipitation, temperature). A few more metrics need to be added (e.g., with climatology as a reference forecast) and the capability to compute aggregated or conditional statistics (using the different attributes used to sort forecasts) needs to be enhanced. An estimation of confidence intervals needs to be added (crucial to deal with small sample sizes). The application should also ingest hindcasts/re-forecasts to support the error analysis work.
3. **Displaying metrics:** Some graphical capability is provided by the IVP GUI, although it needs to be enhanced to provide standardized verification plots and offer more flexibility.
4. **Disseminating the metrics and data:** National verification program provides very limited dissemination of a few metrics; data is not readily available to the public.
5. **Real-time access to metrics:** Not yet available.
6. **Error analysis:** Limited work has been done so far (e.g., archiving stage forecast with and without QPF). No deterministic hindcasting capability has been developed for a comprehensive error analysis work. Work with some RFCs has only started to define raw model forecasts and analyze the performance of the model isolated from forecasters. Various reference and control forecasts should be used, including persistence, climatology, raw model forecast, and simulated hydrologic forecasts (from perfect input). A path for proper error analysis of deterministic forecasts should be identified.
7. **Performance measure tracking:** Not yet available. It is necessary first to compute verification statistics on large samples and numerous forecast points to analyze which performance measures could be used in the future.

FY07

- Define all the data requirements and hardware to archive information for forecast verification and hindcasting purposes for deterministic forecast (including forecast attributes to stratify forecast when computing verification statistics). Determine a plan for developing automated archiving procedures (Item 1). To be done in conjunction with ensemble activities.
- Compute additional metrics as defined by the National Weather Service River Forecast Verification Plan (see Table 2) and develop corresponding graphics in the existing verification software (currently worked on). It should include confidence intervals for most metrics. Add the capability to verify more data types (flow, precipitation, temperature) (currently worked on). Enhance the capability to compute aggregated or conditional statistics (using the different attributes used to sort forecasts). Develop the capability to ingest hindcasts/re-

forecasts to support the error analysis work. Identify key standardized graphic plots to display verification results (Items 2, 3

- Create a plan for disseminating the verification results and data; and create a plan to provide real-time access (Items 4, 5). To be done in conjunction with ensemble activities.
- Identify a path for proper error analysis of deterministic forecasts. Develop a deterministic hindcaster to generate retrospective forecasts from various forecast scenarios, including the use of different meteorological inputs and different sets of model parameters. Establish a definition of the raw model to be accepted by all the RFCs, develop a prototype to allow RFC forecasters to define the raw model, generate and archive raw model forecasts (Item 6).

Ensemble Verification Task Summary

Task	FY06	FY07	FY08	FY09	FY10	FY11
1. Archive						
2. Compute Metrics						
3. Display Metrics						
4. Dissemination						
5. Real Time Access						
6. Error Analysis						
7. Performance Error Tracking						

FY06

1. **Data archiving:** Some capability exists, but must be evaluated to determine if it is sufficient for verification and hindcasting; this data requirement should be done for ensemble prediction in parallel with deterministic forecasting.
2. **Computing metrics:** The Ensemble Verification Program (EVP) prototype has been developed to verify precipitation, temperature, and discharge (including the computation of aggregate statistics for a group of basins). Some probabilistic verification statistics need to be added (such as continuous RPS, Rank Histogram, discrimination measures) as well as additional functionalities to sort forecasts, compute aggregate or conditional statistics, and to estimate confidence intervals (crucial to deal with small sample sizes).
3. **Displaying metrics:** A display capability has been developed based on R scripts to generate jpeg files using the output verification results from EVP. It works for precipitation, temperature, and discharge. It needs to be enhanced to offer more flexibility and to be more user-friendly.
4. **Disseminating the metrics and data:** Not yet available.
5. **Real-time access to metrics:** Not yet available.

6. **Error analysis:** Some capabilities have been developed along with EVP. The EVP prototype could use observed flow values as well as simulated flow values to compute verification statistics and separate the errors from meteorological inputs and from hydrologic models. It also includes deterministic verification statistics for the ensemble mean, persistence, and climatology forecasts. An Ensemble Hindcaster prototype has been developed for both meteorological (precipitation and temperature) and hydrologic forecasts, to produce hydrologic forecasts based on several meteorological forecast scenarios, which are then evaluated by EVP. The robustness of the Ensemble Hindcaster prototype software needs to be improved.
7. **Performance measure tracking:** Not yet available.

FY07

- Evaluate the sufficiency of existing archiving capabilities for ensemble forecasting through interaction with the RFCs. Propose a plan for new needed archiving capabilities (Item 1). To be done in conjunction with deterministic activities.
- Improve the ensemble verification capabilities: enhance the EVP prototype to implement additional statistics (such as continuous RPS, Rank Histogram, discrimination measures) and confidence intervals; enhance the verification graphic capabilities and identify key standardized graphic plots to display verification results. Release and support the experimental version of EVP to the field (Items 2, 3).
- Identify a path for proper error analysis of ensemble forecasts, including the conversion of probabilistic forecasts into deterministic forecasts and comparison with deterministic forecasts. Enhance the Ensemble Hindcaster to analyze the performance of hydrometeorological and hydrologic forecasts based on various forecasting scenarios. Improve the user-friendliness, and release and support the experimental version of the Ensemble Hindcaster (Item 6).
- Create a plan for disseminating the verification results and data; and create a plan to provide real-time access (Items 4, 5). To be done in conjunction with deterministic activities.

Grid Verification Task Summary

Task	FY06	FY07	FY08	FY09	FY10	FY11
1. Archive						
2. Compute Metrics						
3. Display Metrics						
4. Dissemination						
5. Real Time Access						
6. Error Analysis						
7. Performance Error Tracking						

FY06

1. **Data archiving:** Not yet available.
2. **Computing metrics:** Not yet available.
3. **Displaying metrics:** Not yet available.
4. **Disseminating the metrics and data:** Not yet available.
5. **Real-time access to metrics:** Not yet available.
6. **Error analysis:** Not yet available.
7. **Performance measure tracking:** Not yet available.

FY07

- Propose a plan for archiving the data needed for grid forecast verification (Item 1).
- Research metrics to be used for grid forecast verification Q4

References

Brier, G.W., and R.A. Allen, 1951: Verification of Weather Forecasts, Compendium of Meteorology, T.F. Malone, Ed., Amer. Meteor. Soc., 841-848.

Franz, K. J., and S. Sorooshian, 2002: Verification of National Weather Service Probabilistic Hydrologic Forecasts, University of Arizona, report prepared for the National Weather Service.

Joliffe, I.T. and D. B. Stephenson, (ed), 2003: Forecast Verification, , A Practitioners Guide in Atmospheric Sciences, Wiley, West Sussex, England.

Loucks, D.P., J.R. Stedinger and D. Haith, 1981: Water Resources Systems Planning and Analysis, Prentice-Hall, Eaglewood Cliffs, New Jersey.

National Weather Service (NWS), 2006: Interactive Verification Program User's Manual, Silver Spring, MD.

R Development Core Team, 2005: R, A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Walpole, R.E. and R.H. Meyers, 1985: Probability and Statistics for Engineers and Scientists, Macmillan, New York, New York.

Welles, E., N. Cajina, and H. Herr, 2002: Verification of National Weather Service River Stage Forecasts, Second Federal Interagency Hydrologic Modeling Conference, Las Vegas, NV, Jul. 28-Aug. 1, 2002.

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

Wick, G., 2003: Evaluation of Potential Forecast Accuracy Performance Measures for the Advanced Hydrologic Prediction Service, NOAA/NWS/OHD Internal Report.

Wilks, D.S., 1995: Statistical Methods in Atmospheric Sciences, Academic Press, San Diego, California.

Wilson, L.J., 2003: Strategies for the verification of ensemble weather element forecasts,www.emc.ncep.noaa.gov/seminars/presentations/2003/Wilson.presNCEPNov03.ppt, Meteorological Service of Canada, Montreal, Quebec.

Wilson, L.J. and W.R. Burrows, 2004: Spatial verification using the Relative Operating Characteristic curve, AMS, 17th Conference on Probability and Statistics in the Atmospheric Sciences.

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

Appendix 1: Forecast Error Sources

The following is a list of potential sources of error that are to be considered during the river forecasting process**.** This compilation may not be complete as other unique sources of error may exist from RFC to RFC and certainly from differing areas of the country. Additions may be needed in the future.

Data Processing and Quality Control

1. Observed data
   - Available/estimated data
     - Precipitation
       - Type and distribution over basin – rain/snow/freezing rain/sleet
       - Rate (intensity) and distribution over a basin
     - Temperature
     - Stage/Flow
     - Snow depth
     - Snow water equivalent
     - River ice
     - Dew point
     - Wind speed
     - Sky cover: limited or no sunshine measurements
     - Potential evaporation
     - Groundwater
     - Soil moisture
     - Others
   - Source of measurements:
     - Gages:
       - Maintenance and outages
       - Inaccuracies
         - Example for precipitation: inaccuracies in measurements due to snow, sleet, freezing rain and/or wind
       - Density of gage network both within RFC boundaries and between different RFCs
     - Radar:
       - WSR-88D coverage
       - WSR-88D precipitation estimates
         - Z/R relationships
         - Hail contamination
       - Bright banding
     - Satellite
     - Mixture of measurements from different sources
     - Others

2. Forecast data & Climate Outlooks
   - Forecast variable:
     - Precipitation
     - Temperature
     - Stage/Flow
     - Wind Speed
     - Freezing level
     - Others
   - Forecast source

3. Rating Curves
   - Stage/Flow relationship

4. Reservoir outflows/releases including locks and dams both current and future conditions
   - Natural flow vs. regulated flow
   - Other diversions

Modeling

1. Model parameters
   - Types of model:
     - Hydrologic & hydraulic models
     - Data assimilation model
     - Post-processor model
   - Model Calibration:
     - Historical data analysis: selection, quality control, modifications
     - Manual vs. automatic calibration
     - Calibrator's experience and/or ability
   - Other parameter concerns
     - Rainfall/Runoff
     - Snow accumulation/ablation
     - Consumptive Use
     - Routing
     - Dynamic Routing
     - 6 – hour time step limitations
     - Operational Implementaion

3. Model Structure
   - Spatial and temporal scales:
     - Raw data vs. ingested or output data
       - Example: lumped model assumes uniform conditions over up to hundreds of square miles for 6 hours
       - Disaggregation process to interpolate 6-hour values from 24-hour data
         - to estimate 6-hour mean values from maximum/minimum daily temperature values
         - to estimate 6-hour flow values from mean daily flows
   - Other limitations/deficiencies
     - Modeled processes vs. actual processes
     - Changes over time (land use, river system, etc.)

2. Initial Model States
   - Manual vs. automatic state updating
     - Adjustments of parameters, model states, input data
     - Data assimilation model

Forecaster Interaction

1. Forecaster interpretation
   - Quality control of input data
   - Reality check of forecasts
   - Run-time mods
     - Modification of input data (observed and forecast), model parameters, initial model states
   - Rules of thumb

2. Forecaster experience
   - New/updated technology or capability
   - New/updated input (observed and forecast)
   - New/updated model (parameters, model states, structure)

Appendix 2: Definition of Metrics for the National Baseline Verification System

The National Baseline Verification System (NBVS) should provide the following verification metrics for *administrative* verification purposes. Definitions for the statistics for the NBVS are listed below:

Probability of detection (POD) – Percentage of (categorical) events forecast correctly.

False Alarm Ration (FAR) – Percentage of (categorical) forecast events that did not verify.

Lead Time of Detection (LTD) – The average lead time of all forecasts that fall into the correct observed category.

Brier Score (BS) - The mean squared 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.

Ranked Probability Score (RPS) – The mean squared 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.

Root Mean Square Error (RMSE) – The square root of the average of the squared differences between forecasts and observations.

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

Mean Absolute Error (MAE) – The average of the absolute value of the differences between forecasts and observations.

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

Mean Error (ME) – The average difference between forecasts and observations.

Bias (%) – The ME expressed as a percentage of the mean observation.

Correlation Coefficient – A measure of the linear association between forecasts and observations.

Skill Score – In general, skill scores are the percentage difference between verification scores for two sets of forecasts (e.g., operational forecasts and climatology).

$$SS = 1 - \frac{Score(\ forecastset1)}{Score(\ forecastset2)}$$

Root Mean Squared Error Skill Score (SS-RMSE) – A skill score based on RMSE values. The recommended reference forecasts are persistence and climatology.

Brier Skill Score (BSS) – A skill score based on BS values. The recommended reference forecasts are persistence and climatology.

Ranked Probability Skill Score (RPSS) – A skill score based on RPS values. The recommended 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.)

Appendix 3: Example Calculations for the National Baseline Verification System (NBVS)

In this section, sample calculations using a small, simplistic data set of forecast and observations are presented for the NBVS. The sample data set is shown below:

Year	Obs	Deterministic Forecast	Ensemble (Probabilistic) Forecast			
			E1	E2	E3	E4
1	112	72	42	74	82	90
2	**206**	165	65	143	223	227
3	**301**	218	82	192	295	300
4	**516**	417	211	397	514	544
5	**348**	285	142	291	349	356
6	98	275	114	277	351	356
7	156	170	98	170	204	205
8	**245**	176	69	169	229	236
9	**233**	213	94	219	267	270
10	**248**	182	59	175	244	250
11	**227**	188	108	089	227	228
12	167	136	94	135	156	158

Table 1: Sample forecast / observation pairs for a peak flow forecast on July 1 over the years 1-12 for a mythical forecast point. All values are in cfs. Flood observations are shown in bold.

The forecasts and observations are shown for a peak flow forecast made on July 1 for a hypothetical forecast point. The flood flow for this forecast point is 300 cfs. A forecast ensemble with four ensemble members was made. The deterministic forecast was derived by taking the mean value of the ensemble.

1. NBVS deterministic metrics

A. Categorical Metrics

For the categorical verification metrics, a contingency table is constructed relative to the flood flow of 300 cfs using the deterministic forecasts and observations from table 1:

	# Observed flood	# Observed no flood
# Forecast flood	(a) 4	(b) 1
# Forecast no flood	(c) 4	(d) 3

The Probability Of Detection (POD) is the number of events both forecast and observed divided by the number of observed events:

$$POD = \frac{a}{a+c} = \frac{4}{4+4} = 0.5 = 50\%$$

The False Alarm Ratio (FAR) is the number of events falsely forecast to occur divided by the total number of “non-events”:

$$FAR = \frac{b}{a+b} = \frac{1}{4+1} = 0.2 = 20\%$$

*Lead time - Can not be calculated from this sample size, but is defined as the time of issuance of the forecast until the time the event occurs e.g. flood stage.

B. Error Metrics

The deterministic error metrics are derived from the difference between the forecast and the observation pairs:

Year	Obs	Deterministic Forecast	Error (Forecast – Obs)	Absolute Error	Squared Error
1	112	72	-40	40	1600
2	**206**	165	-41	41	1681
3	**301**	218	-83	83	6889
4	**516**	417	-99	99	9801
5	**348**	285	-63	63	3969
6	98	275	177	177	31329
7	156	170	14	14	196
8	**245**	176	-69	69	4761
9	**233**	213	-20	20	400
10	**248**	182	-66	66	4356
11	**227**	188	-39	39	1521
12	167	136	-31	31	961
			ME = 30	MAE = 61.8	RMSE = 75

The mean error (ME) is the average of the differences between forecast / observation pairs. The mean absolute error (MAE) is the mean of the absolute error. The RMSE is the square root of the mean of the squared differences. The bias in percentage terms is the ME divided by the mean observation or -12.6%.

C. Skill Score

From the definition of Skill Score (SS): SS = 1 - MSE/MSE$_{climatology}$. From the sample data set table (above), using ***R***, we have:

```
> obs<-c(112,206,301,516,348,98,156,245,233,248,227,167)
> forecast<-c(72,165,218,417,285,275,170,176,213,182,188,136)
> climate<-c(238,238,238,238,238,238,238,238,238,238,238,238)
```

Using the ***R*** *verification* package:

```
> B<-verify(obs, forecast,frcst.type = "cont", obs.type = "cont")
> summary(B)

The forecasts are continuous, the observations are continous.
Sample baseline calcluated from observations.
MAE               =  61.83
ME                =    -30
MSE               =   5622
MSE - baseline    =  1.183e+04
MSE - persistence =  1.55e+04
```

```
SS  - baseline    =   6205
```

The MSE = 5622, the $MSE_{baseline} = MSE_{climatology}$ = 1.183e+04, so SS = 1 – 5622/1.183e+04 = 0.4752. Consequently, since SS is greater than zero, the forecast shows skill with respect to the baseline or climatology forecast.

D. Sample Size

In this case, the sample size is 12 forecast / observation pairs. This is a relatively small sample size. Therefore caution should be exercised when drawing conclusions from this particular example.

Please refer to Appendix 4 for another example of the effect of sample size.

2. NBVS probabilistic metrics

A. Categorical Metrics

The Brier Score (BS) calculation is broken down in the table below:

Year	Obs	Ensemble (Probabilistic) Forecast				Flood? (o)	Forecast Probability (Flood) (y)	Brier Score (y-o)^2
		E1	E2	E3	E4			
1	112	42	74	82	90	0	0	0
2	**206**	65	143	223	227	1	0.5	0.25
3	**301**	82	192	295	300	1	0.5	0.25
4	**516**	211	397	514	544	1	0.75	0.06
5	**348**	142	291	349	356	1	0.75	0.06
6	98	114	277	351	356	0	0.75	0.56
7	156	98	170	204	205	0	0.5	0.25
8	**245**	69	169	229	236	1	0.5	0.25
9	**233**	94	219	267	270	1	0.75	0.06
10	**248**	59	175	244	250	1	0.5	0.25
11	**227**	108	189	227	228	1	0.5	0.25
12	167	94	135	156	158	0	0	0

First, the observations and forecasts are characterized according to flood criteria (200 cfs). Observations will either be above flood (o=1) or below flood (o=0). Forecasts may have some percentage above or below. For example, the forecast for year 2 has two ensemble members below flood stage and two above. Therefore the probability of flood is 0.5 or 50%. Next, the squared difference between the forecasts and observations are calculated. Finally, the BS is taken as the average of the squared differences. In this case, the BS is 0.187.

The Ranked Probability Score (RPS) is a multi-category extension of the BS. For simplicity, the categories for the sample calculation will be: 100,200,300, and 400 cfs. The RPS calculation is shown in the table below:

Year	Prob(observed flow < …)				Prob(forecast flow < …)				RPS
	100	200	300	400	100	200	300	400	Σ(yi-oi)^2
1	0	1	1	1	1	1	1	1	1
2	0	0	1	1	0.25	0.5	1	1	0.31
3	0	0	0	1	0.25	0.5	1	1	1.31
4	0	0	0	0	0	0	0.25	0.5	0.31
5	0	0	0	1	0	0.25	0.5	1	0.31
6	1	1	1	1	0	0.25	0.5	1	1.81
7	0	1	1	1	0.25	0.5	1	1	0.31
8	0	0	1	1	0.25	0.5	1	1	0.31
9	0	0	1	1	0.25	0.25	1	1	0.13
10	0	0	1	1	0.25	0.5	1	1	0.31
11	0	0	1	1	0	0.5	1	1	0.25
12	0	1	1	1	0.25	1	1	1	0.06

First, the observations and forecasts are characterized according to probability of exceedence of each category threshold. Each threshold is treated like the single threshold used for the BS. Next, the sum of the squared differences is calculated and summed for each threshold. Mathematically this is given by:

$$RPS = \sum_{i=bin\#1}^{bin\#n} [P(\,forecast < i) - P(observed < i)]^2$$

Finally, a mean RPS is computed across all forecasts. In this case, the mean RPS is 0.535. This calculation allows comparisons between forecasts. In this example, year 6 was the worst forecast (RPS = 1.81) and year 12 (RPS = 0.06) was the best forecast.

B. Skill Score

For the sake of this example, assume the BS for climatology forecasts was 0.25 and the RPS for climatology forecasts was 1.0. Therefore, the BSS is given by:

$$BSS = 1 - \frac{BS}{BS_{ref}} = 1 - \frac{0.187}{0.25} = 0.25 = 25\%$$

This means the forecast was 25% better than using climatology as a forecast using the two categories in the BS calculation.

The RPSS is given by:

$$RPSS = 1 - \frac{RPS}{RPS_{ref}} = 1 - \frac{0.535}{1.0} = 0.465 = 46.5\%$$

This means the forecast was 46.5% better than using climatology as a forecast with multiple categories.

C. Sample Size

The sample size is the same as the deterministic example, 12 forecast / observation pairs. However, in this case there is the added information that the forecast ensemble has four members.