NWS Hydrology Forecast Verification Team

Interim Report
January 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 (NCRFC),
Rob Shedd and Tom Econopouly (NERFC),
Joe Intermill (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

Table of Content

Introduction

The NWS Hydrology Forecast Verification Team was chartered in July 2007 with the following team charter.

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 determine actions 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 March 31, 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, includes 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. The team had 12 teleconferences 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. The team then met again at the second RFC Verification Workshop on November 18-20, 2008.

This report includes the data archiving requirements report delivered to this team (and the IWT Archive Team) in May 2008 (the results from the archiving survey filled by the 13 RFCs are given in the Appendix), the 13 RFC verification case studies, and the recommendations and actions from the second verification workshop to prepare the final team report and propose standard verification strategies.

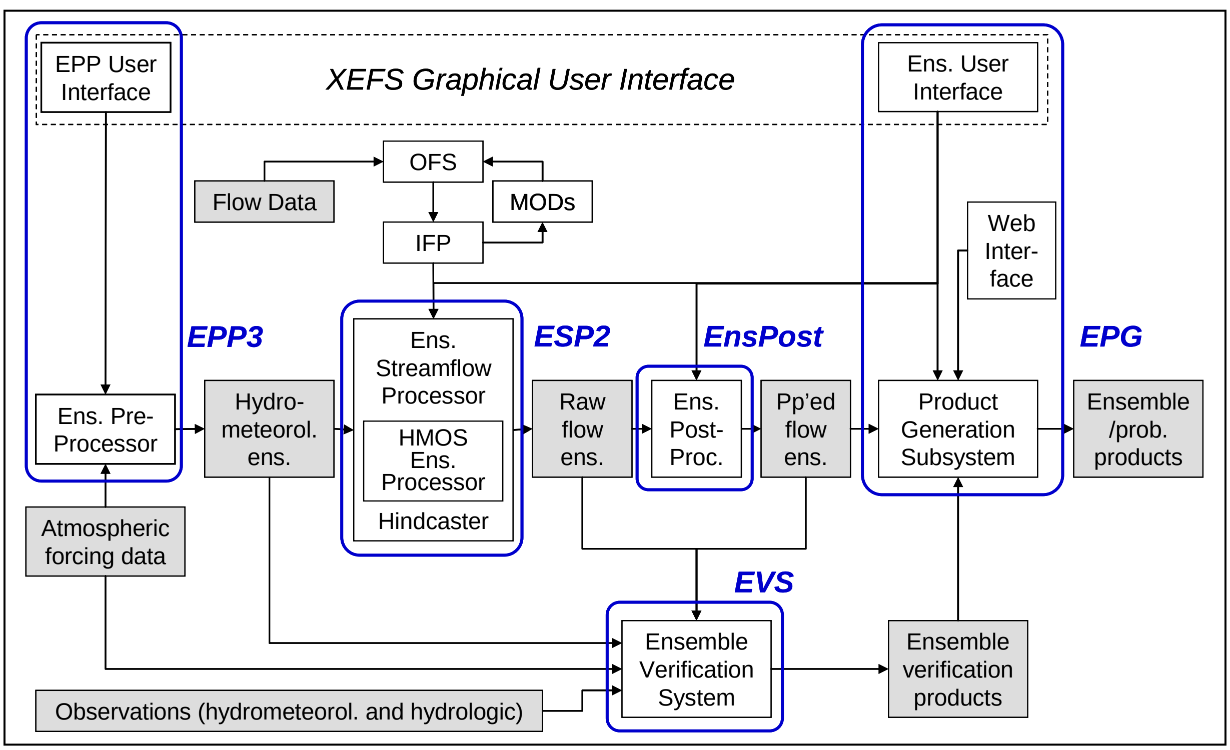
Data Archiving Requirements for Forecasting and Verification Purposes
May 2008

<u>Goal</u>: define the data archiving requirements to perform single-valued and probabilistic forecasting, as well as hindcasting and verification, and to disseminate all the necessary information on products and services to partners and users. These data requirements will support CHPS and XEFS and will be used by the IWT Archive Team led by Julie Meyer.

The users are:

- RFC forecasters
- Scientists from NWS and the wider hydrologic community
- Software engineers from NWS and the wider hydrologic community
- Hydrology Program Managers
- General public with a wide range of users and partners, from the general public to sophisticated users ingesting forecast products into their own risk decision system.

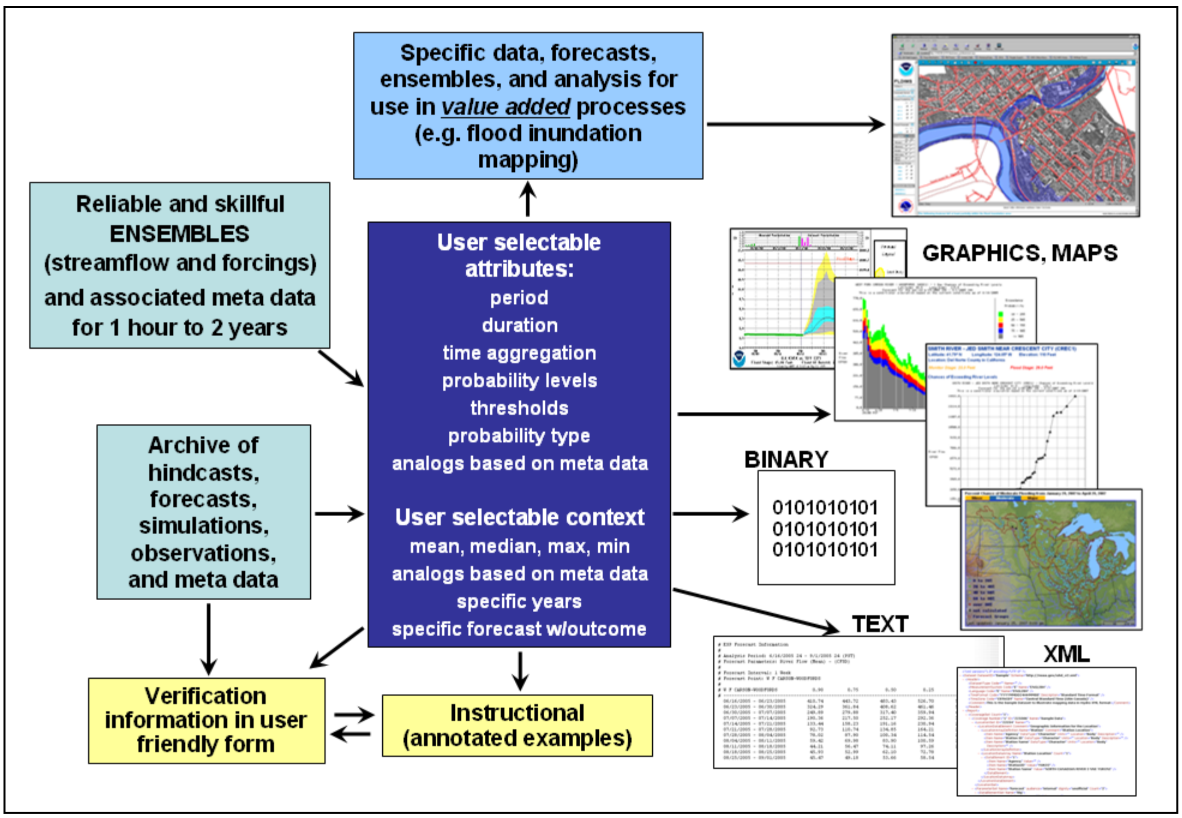
XEFS is the prototype currently developed to produce both single-valued and ensemble forecasts within CHPS. The different XEFS components and data flow are shown in the figure below, as described in the XEFS Design and Gap Analysis Report (NWS, 2007), which is available at http://www.nws.noaa.gov/oh/rfcdev/docs/XEFS_design_gap_analysis_report_final.pdf.



The acronyms are:

- EPP3, the Ensemble Pre-processor 3, which combines the EPP2 and GFS sub-systems
- ESP2, the next generation of the Ensemble Streamflow Prediction technique
- EnsPost, the Ensemble Post-Processor
- EPG, the Ensemble Product Generation system
- EVS, the Ensemble Verification System, to become the National Baseline Verification System (NBVS) in the future.

Below is a description of products and information to be made available for the XEFS users, as described in the XEFS Design and Gap Analysis Report (NWS, 2007).



Data to archive

This section describes the different data sets to be archived with the archiving priorities for the near future, especially for running the XEFS prototype in the coming 3 years at all RFCs.

Note that it is crucial to archive all data and forecasts that require forecaster intervention since any manual process cannot be reproduced retroactively. For automated processes, all the information needed to retroactively generate the output products need to be archived, not necessarily the output products themselves.

1) Observations:

- raw, quality controlled (using different techniques, manual and/or automated), finalized;
- for precipitation (type and rate, with distribution over basin), temperature, potential evaporation, freezing levels, groundwater, soil moisture, snow depth, snow water equivalent, river ice, dew point, wind speed, sky cover, streamflow, stage, reservoir outflows/releases;
- from different sources of measurements (gages, radar, satellite, mixture of measurements, etc.);
- relative to point/area or grid

Archiving priority:

- Mean Areal Precipitation (MAP), Mean Areal Temperature (MAT), potentially other forcing inputs (such as potential evaporation and freezing levels) at 6-hour or 24-hour time step (or less if available) for all forecast points for model calibration (including Ensemble Preprocessor), forecasting/hindcasting, verification, and dissemination
- Streamflow or stage at 6-hour or 24-hour time step (or less if available) for all forecast points for model calibration (including hydrologic and hydraulic models, Ensemble Postprocessor), forecasting/hindcasting, verification, and dissemination; if necessary, reservoir outflows/releases

2) Forecasts/Hindcasts and Simulations:

- raw, operational, and experimental from different sources or models;
- for precipitation, temperature, potential evaporation, soil moisture, snow, wind speed, freezing levels, streamflow, stage;
- single-valued or probabilistic;
- relative to point/area or grid

Archiving priority:

- Operational single-valued forcing input forecasts at 6-hour or 24-hour (or less if available) time step for all forecast points for model calibration (including Ensemble Preprocessor), ensemble forecasting/hindcasting, verification, and dissemination: Forecast Mean Areal Precipitation (FMAP), Forecast Mean Areal Temperature (FMAT), other forcing inputs (such as Potential Evaporation), other single-valued forcing inputs such as GEFS ensemble means for precipitation and temperature

- Probabilistic forcing input forecasts at 6-hour or 24-hour time step (or less if available) for all forecast points for model calibration (including Ensemble Preprocessor), verification, and dissemination: PQPF and PQTF (and potentially other forcing inputs) from the different existing ensemble methodologies (Ensemble Preprocessor including RFC sub-system and GFS sub-system, generic ensembles for weather and climate from NCEP or other sources – GEFS, SREF, CFS, CPC outlook)

- Single-valued forecasts and simulations of streamflow or stage (the rating curves need to be archived to convert streamflow to stage and vice versa) at 6-hour or 24-hour time step (or less if available) for all forecast points for model calibration (including Ensemble Postprocessor, HMOS), ensemble forecasting/hindcasting, verification, and dissemination; these should include raw model forecasts (without any manual intervention) and finalized MOD-ed forecasts as used in the operations

- Probabilistic and statistical forecasts of streamflow or stage at 6-hour or 24-hour time step (or less if available)for all forecast points for model calibration (Postprocessor), verification, and dissemination

3) Models set-up:

- definition of forecast points/areas,

- definition of models suite,

- definition of spatial and temporal scales (spatial resolution, temporal resolution/time step, forecast horizon/lead time, variable type – instantaneous or aggregation statistic)

4) Models parameters:

- Parameter sets for all models (hydrologic, hydraulic, routing, rating curves, data assimilation (DA), preprocessor, postprocessor, etc.)

Archiving priority:

- Past and current rating curves and shifts for all forecast points to convert streamflow to stage and vice versa for model calibration and single-valued and ensemble forecasting/hindcasting

- Past and current parameter sets for current models, including the parameters for Ensemble Preprocessor, ESP and Ensemble Postprocessor, for model calibration, and single-valued and ensemble forecasting/hindcasting

5) Initial conditions:

- raw, operational, and experimental;
- from different processes (forecaster input, various run-time MODs, DA)

Archiving priority:

- Raw (without any forecaster intervention) and operational (with run-time MODs and/or DA technique) sets of initial conditions for all forecast points for model calibration and single-valued and ensemble forecasting/hindcasting

6) Verification information:

- forecast service (logistical) verification to evaluate the quality of delivered forecast services in terms of the service efficiency and forecast usability (e.g., number of forecasts locations, service type, frequency, forecast timeliness);
- forecast diagnostic verification to assess the quality of past forecasts given certain conditions (time, forecast/observed value, event, methodology...)
- forecast real-time (prognostic) verification to evaluate the quality of operational forecasts in real-time using observations associated with historical forecasts analogous to the real-time forecast, and potentially correct forecast errors detected in the prognostic verification results.
- verification products (numerical results, graphics, maps, formatted reports)

Archiving priority:

- Forecast services data in the IHFS-DB (to become available in the National River Location Database) for current forecast points
- All information on forecast points to do verification analysis with IVP (for single-valued forecasts), EVS (for ensemble forecasts), or other forecast verification application for forcing input forecasts/hindcasts and hydrologic forecasts/hindcasts to run/re-run the software and generate verification metrics; information should include observation files, forecast/hindcast files, and verification parameters as required by current verification application

7) Metadata for all archived data (observations, forecasts, model parameters, model initial conditions, verification):

   - date and time (including time system), location, flow/stage critical thresholds, upstream/downstream segments
   - time step, scale (instantaneous or aggregated over period, and aggregation statistic if aggregated (e.g. mean, accumulation etc.))
   - units of measurement
   - how data/product was generated

Tasks to be performed by users:

1) importing and storing data and metadata, with monitoring of archiving process
2) display and quality control of archived data
3) exporting archived data
4) model(s) calibration (parameter estimation and calibration) (manual and automatic), forecast verification, and logistical verification
5) model(s) training
6) real-time forecasting, real-time forecast verification to provide decision-support to forecasters, and logistical verification
7) hindcasting (re-forecasting) to support verification work
8) display and statistical analysis of input and output data (with user selectable attributes and context information)
9) forecast verification with the river forecast verification system
10) generation of user specified products (forecasts, hindcasts, observations, simulations, verification products, metadata), as well as dissemination of data and products

The appendix is the archiving survey results from the 13 RFCs as reported to the NWS Verification team in February 2008, as well as archiving data requirements for the current XEFS prototype.

Formal hardware requirements will be developed by the IWT Archive Team led by Julie Meyer.

Verification Case Studies from the 13 River Forecast Centers
March-November 2008

As part of the NWS Hydrology 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 deterministic 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 to improve their forecasts by doing such post-event and diagnostic verification analysis.

The case studies presented in this report by the 13 RFCs are relative to the following forecast quality analysis:

- CNRFC: operational forecasts of precipitation, temperature and flow during a major flood event for 2 river basins
- NWRFC: forecasts of one extreme event for one basin, including the retrospective analysis of the error sources (using IFP)
- CBRFC: raw model forecasts and simulations of a flood event for 2 basins to evaluate the impact of QPF errors on flow forecasts
- LMRFC: forecasts during Hurricane Katrina in comparison to forecasts of events from five years, for 4 forecast points
- NCRFC: operational forecasts and QPF contingency forecasts during the flooding events in 2008
- MBRFC: operational forecasts of flow and stage for four basins, with persistence as the baseline forecast
- SERFC: operational forecasts for one basin during the last 8 years and proposed analysis of errors in forecast shape and forecast timing
- NERFC: QPF forecasts from various sources (HPC, NDFD, NERFC) and from different individual forecasters for 13 basins
- OHRFC: operational forecasts and 6 sets of experimental forecasts from different scenarios (MODs/no-MODs, no-QPF/HPC-QPF/HAS-QPF) for 7 basins
- APRFC: stage forecasts from two calibration strategies for one basin
- WGRFC: comparison of VAR forecasts with operational forecasts for 3 basins
- ABRFC: ensemble forecasts produced by the HMOS prototype software for 3 basins
- MARFC: ensemble forecasts produced by the Ensemble Preprocessor prototype (EPP2) and ESP for 2 basins

The presentations of these verification case studies are 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_docs.html

CNRFC Verification Case Study
Presented on June 24, 2008 by Alan Takamoto

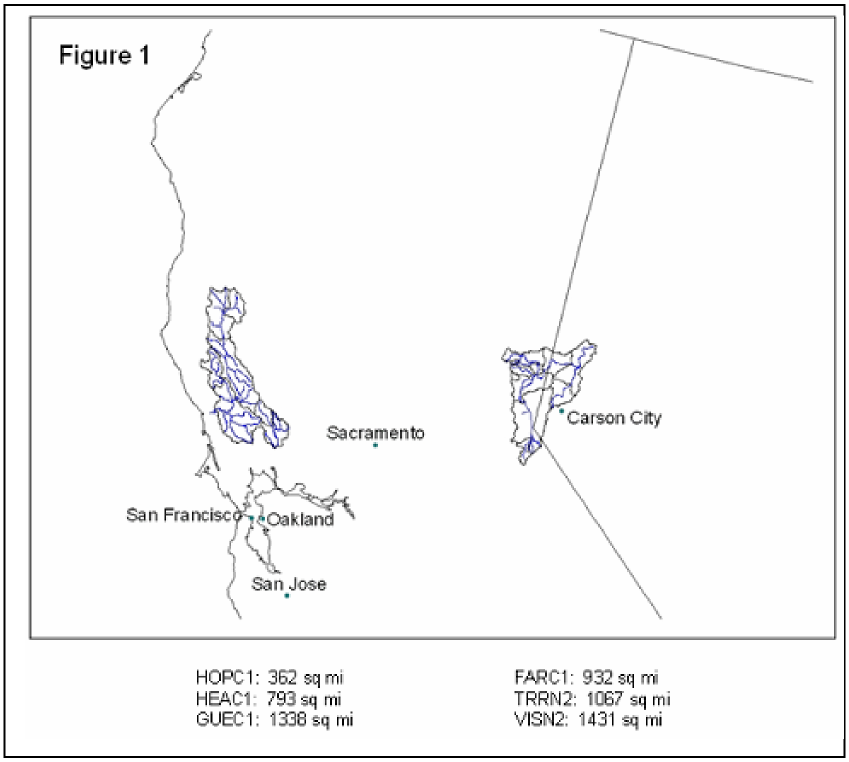
Abstract

This paper describes the implementation of the National Weather Service's Interactive Verification Program (IVP) for two flood events during the New Year's 2006 storm event in California and Northern Nevada. Flooding due to extreme precipitation events were examined for the Russian River basin in Mendocino and Sonoma counties in California and the Truckee River basin whose headwaters start just above Lake Tahoe in California's Sierra Nevada and eventually drains into Pyramid Lake in Nevada. Various statistical display tools are available from IVP, including scatter plots, time series plots, lead time statistics and categorical statistics. Statistical analysis were made for river flow, mean areal precipitation and mean areal temperature forecasts. Initial findings indicate significant under-forecasting of flow in the Truckee basin during the flood event, which was brought out quite effectively on the IVP scatter- and time series plots for flow and mean areal precipitation.
The study presented below does not include all the figures; the complete paper is available online at http://www.nws.noaa.gov/oh/rfcdev/docs/CNRFC_verification_case_study_report_Jan09.pdf

Synopsis

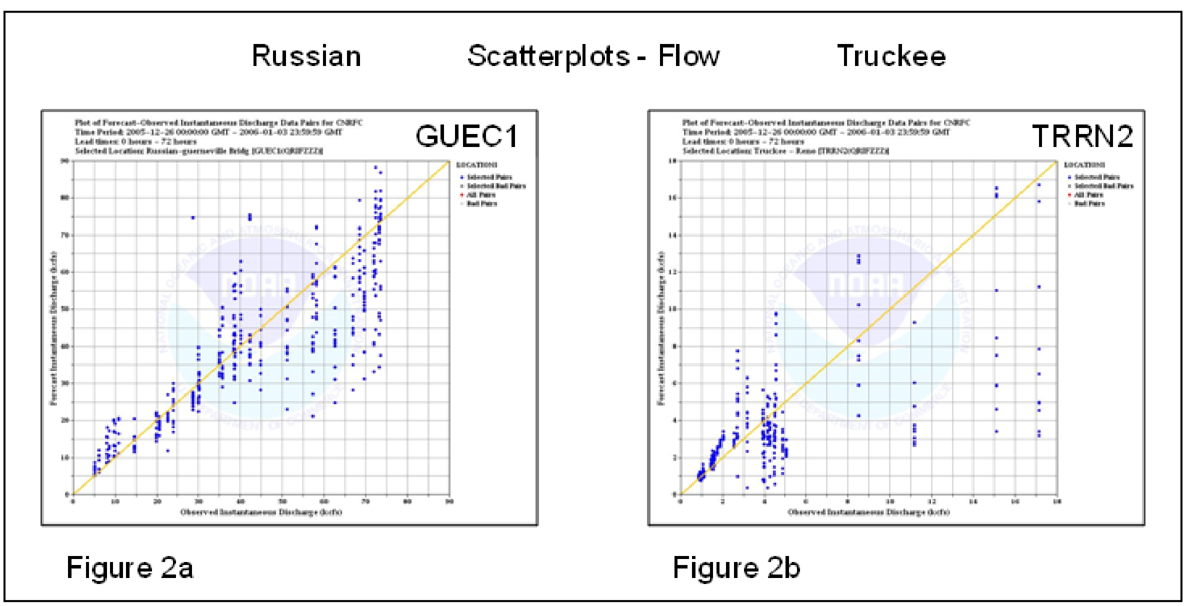
This is an expansion of the case study of the 2006 New Year's Flood event on the Truckee River presented to the National Weather Service, Western Region Verification team on December 12, 2007. This study will also examine the flooding that occurred on the Russian River during the same New Years' event.
A rather productive stormy period occurred during December 26, 2005 through January 3, 2006. Soil moisture conditions increased with each passing storm until a large storm around December 28th and a significantly wetter event on New Year's Eve brought forecast points to above flood stage in California's north coast and Russian and Napa Rivers. Forecast points on the lower Truckee River in Nevada also rose above flood stage. Both the Russian and Truckee River basins experienced significant damage during the New Year's event, with about $108 million in damages to businesses and residences in Sonoma County alone. Major flooding occurred in Guerneville and Healdsburg. Northern Nevada experienced about $17 million in damages from the same event. Much of the commercial business area adjacent to the Truckee River in Sparks, Nevada, was 2 to 4 feet under water.

Study area (Figure 1) - The Russian River basin drains into the Pacific Ocean and is generally subject to orographically enhanced rainfall events. The headwaters of the basin begin in Mendocino County and flow through Sonoma County. The basin size is in excess of 1300 sq mi. Major communities impacted include Guerneville and Healdsburg. The river forecast points studied include the Russian River at Hopland (HOPC1), the Russian River at Healdsburg (HEAC1) and the Russian River at Guerneville (GUEC1). The headwaters of the Truckee River start just above Lake Tahoe and eventually drain into Pyramid Lake in Washoe County, Nevada. The basin is located just east of the crest of the Sierra Nevada and is subject to "spillover" events, where under the proper meteorological conditions, copious precipitation can spill over the lee side of the Sierra crest into the Truckee basin. The basin size is over 1800 sq mi in size. The major towns affected by flooding events include the town of Truckee CA, and downtown Reno and Sparks, Nevada. Forecast points examined include the Truckee River at Farad (FARC1), the Truckee River at Reno (TRRN2) and the Truckee River at Vista (VISN2).



This case study will compare the CNRFC performance of forecasting the Russian and Truckee River during the same major flood event.

Flow Scatter plots – Flow was generally under-forecast on the Russian River, at the highest discharge rates (Figure 2a). Flow was also under-forecast at the higher amounts on the Truckee River at Reno (Figure 2b).



Time Series for Flow - The time series for flow reveals over-forecasting at the upstream stream gage on the Russian River at Hopland (HOPC1) during the first peak on 12/29/2005, under-forecasting during the 12/31 peak and over-forecasting during the small peak around 01/02/2006 (Figure 3a). Further downstream, there was over-forecasting at Guerneville (GUEC1) during the 12/29/2005 peak (see Figure 3c online) and flow was over-forecast during the damaging peak on 01/01/2006 when the river was significantly above Flood Stage. The recession was under-forecast after the 01/01 peak while it was above flood stage. For the Truckee River, the time series for the upstream point at Farad shows a definite under-forecast before the 12/31 peak (Figure 3b), but the river never reached flood stage. It was also under-forecast at Reno (TRRN2) downstream days before the crest on 12/31 (see Figure 3d online). The hydrologist was able to forecast near the actual peak at the Reno forecast point just a few hours before it occurred.

Time Series - Flow

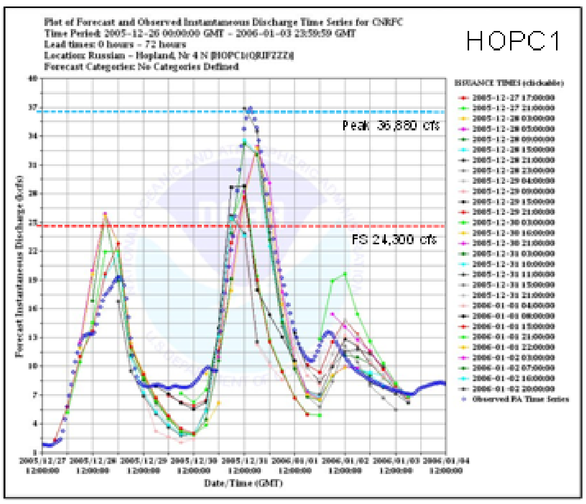


Figure 3a

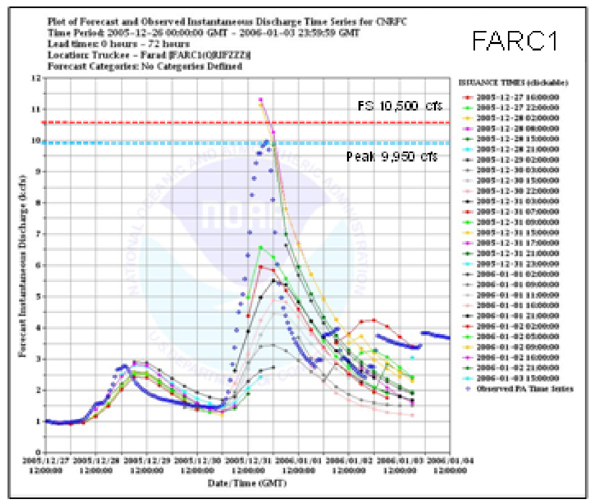
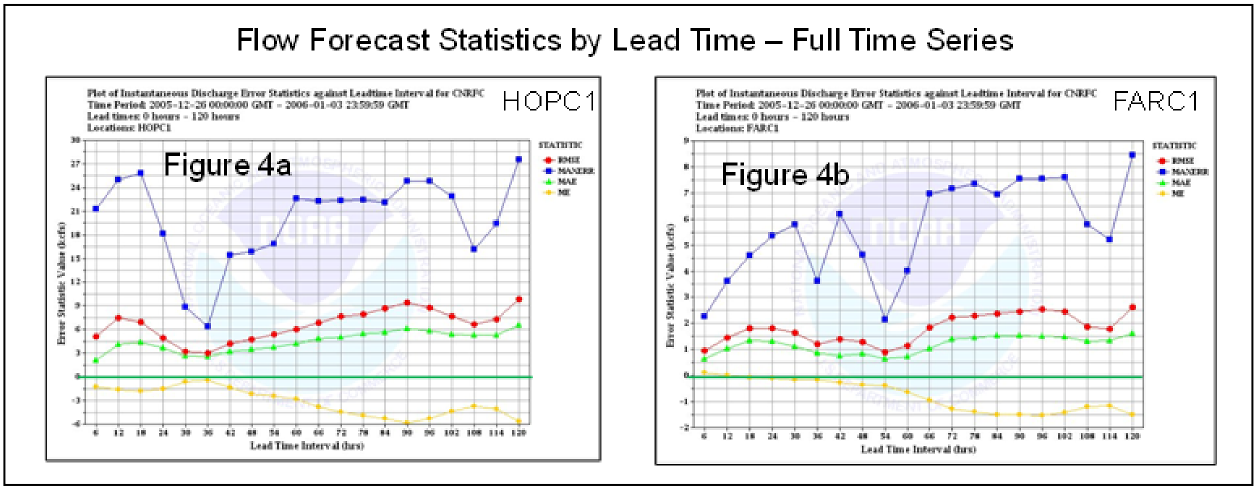
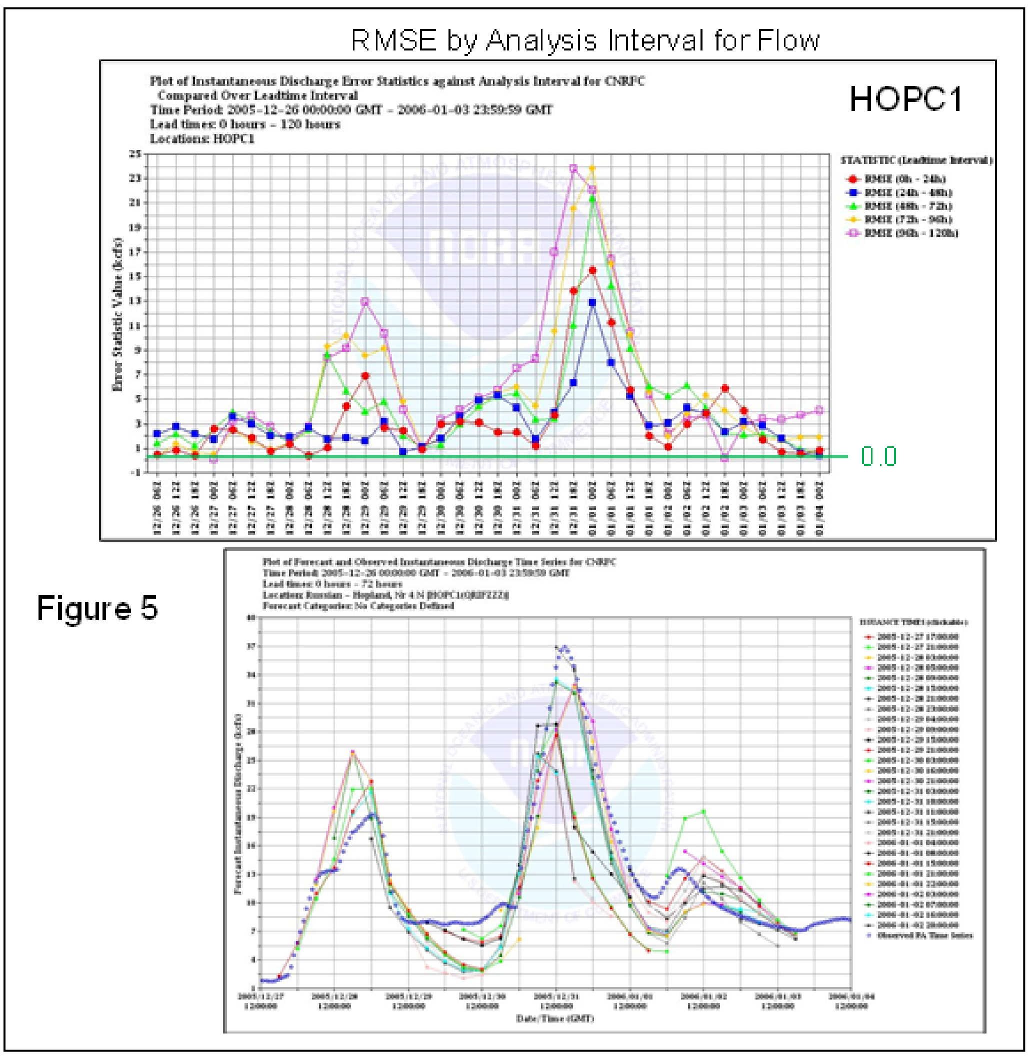


Figure 3b

A plot of instantaneous discharge error statistics was plotted against lead time interval. The Russian basin (Figure 4a below and Figure 4c online) showed generally the highest MAXERR during the earliest lead time (6 to 12 hours) and RMSE rose slowly from 42 to 120 hours. MAXERR increased gradually on the Truckee, with an unexplainable dip between a lead time at 36 hours and 42 to 60 hours. The dip could be a result of the small sample size. Conversely, RMSE and MAE increased gradually on the Russian, especially at GUEC1 (Figure 4c online). There was a gradual increase for the Truckee also (Figure 4b below and Figure 4d online), except that it also dipped at a lead time of 36 hours and 42 to 60 hours. In general, MAXERR and RMSE increased slowly with increasing lead time. Mean Error was negative for both the Russian and Truckee.

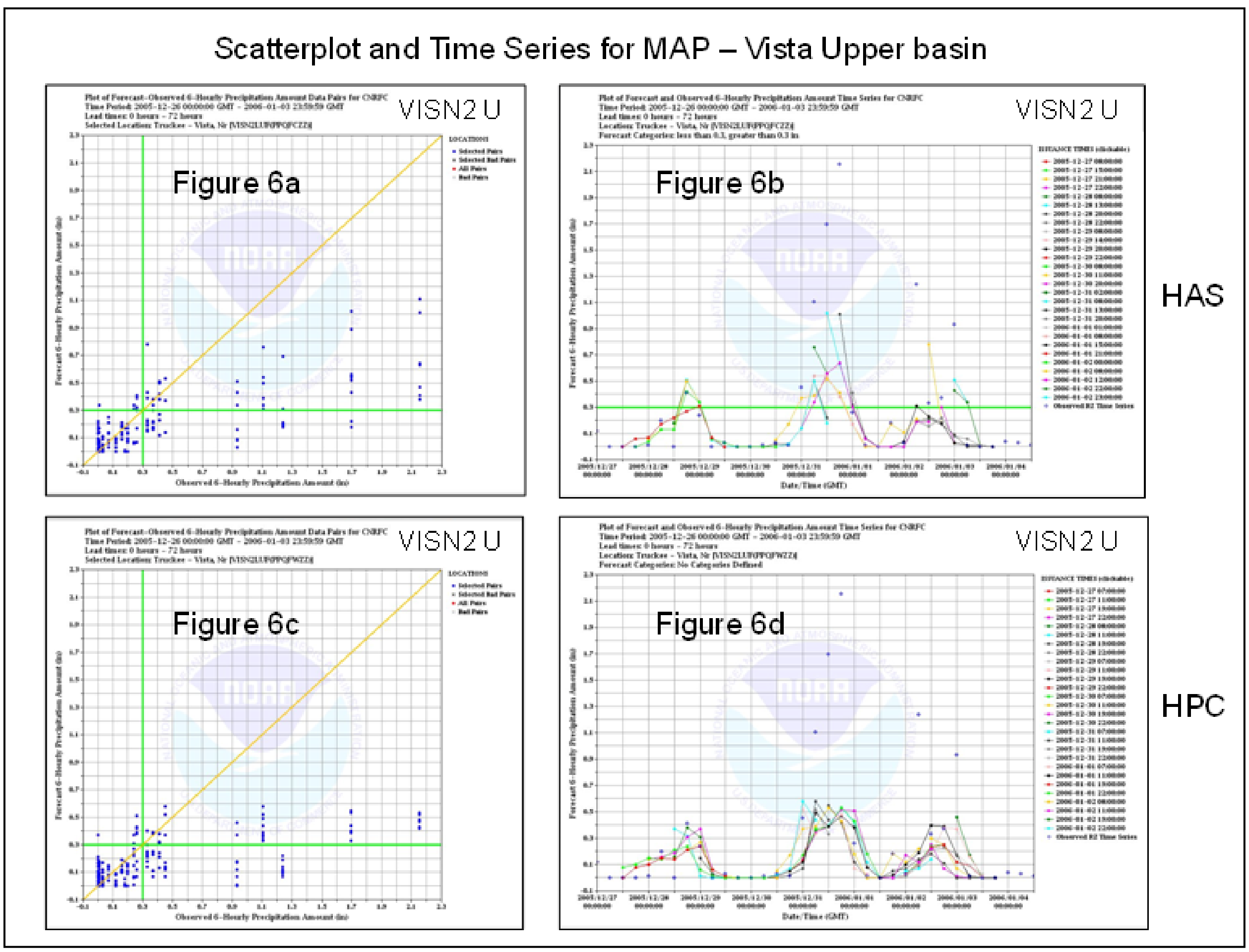


Plot of instantaneous discharge error statistics against analysis interval, computed over lead time interval (Figure 5). The RMSE increased substantially with increasing lead time where the peaks occurred on the hydrograph for both the HOPC1 forecast point on the Russian and FARC1 on the Truckee. This was not the case for the BIAS by Analysis Interval for Flow.



MAP – verification of HAS unit. Quantitative Precipitation Forecasts (QPF) are issued by the CNRFC Hydrometeorological Analysis and Support (HAS) unit. The CNRFC HAS unit is a dedicated unit made up of three meteorologists. Basically, the HAS unit forecasts 6-hourly QPF and freezing level out 0-72 hours (day 1 through 3). For 72-144 hours (days 4 through 6), six-hourly QPF and freezing level forecasts are issued using the Rhea Orographic Aid based on GFS gridded output. The Rhea Orographic Aid enhances the topography as the parcel travels over it to the crest. Generally, most of the effort is put on the day 1 through 3 forecasts by the HAS unit. The 72 to 144 hour QPF forecast was not covered in this case study. For days 1 through 3, the HAS consults the weather models, HPC guidance, collaborate with NWS Weather Forecast Office meteorologists and use other tools such as surface observations and remote data collection available to them on AWIPS and other computers. There is extended coverage by the HAS unit during periods of forecast heavy precipitation. The Mean Areal Precipitation (MAP) observed and forecast data used to generate the statistics for this case study was derived from the HAS and HPC QPF forecasts.

Scatter plots were generated for the upper (VISN2 U) portion of the Truckee Basin at Vista. VISN2 is calibrated in NWSRFS as a split basin with an upper and lower portion delineated by elevation. Note for VISN2 U the relative under-forecast at the higher observed precipitation levels on the scatter plots for both HAS (Figure 6a) and HPC (Figure 6c) and the under-forecast MAP for both on the time series plots (Figures 6b and 6d). It appears that HPC is slightly worse than the HAS forecast for VISN2 at the higher observed values on the scatter plot, but both are well under-forecast.



Moments – MAPs: Figure 7 is a plot of 6-hourly precipitation amount moments (observed mean, forecast mean, observed standard deviation and forecast standard deviation) against analysis interval for GUEC1 and HOPC1. Note that increasing observed and forecast means roughly coincides with the discharge peaks on the time series plot for HOPC1.

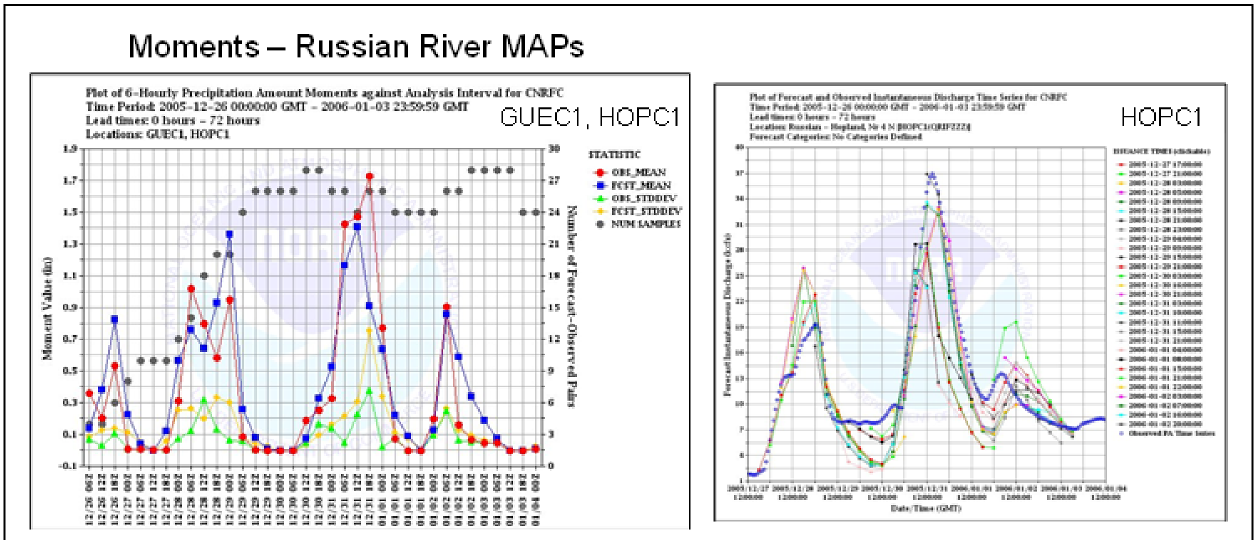
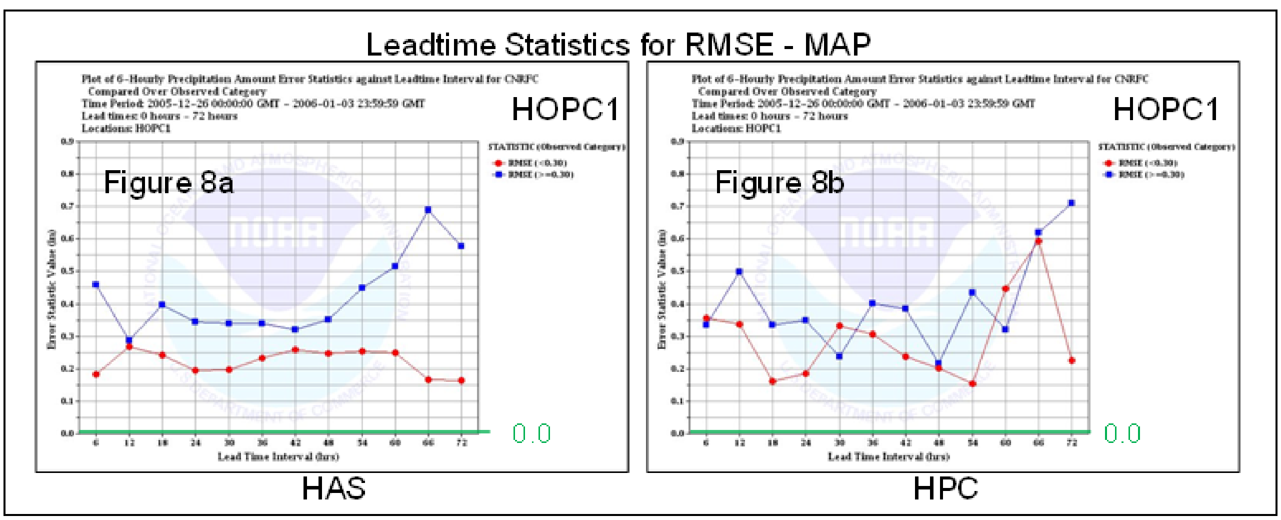


Figure 7

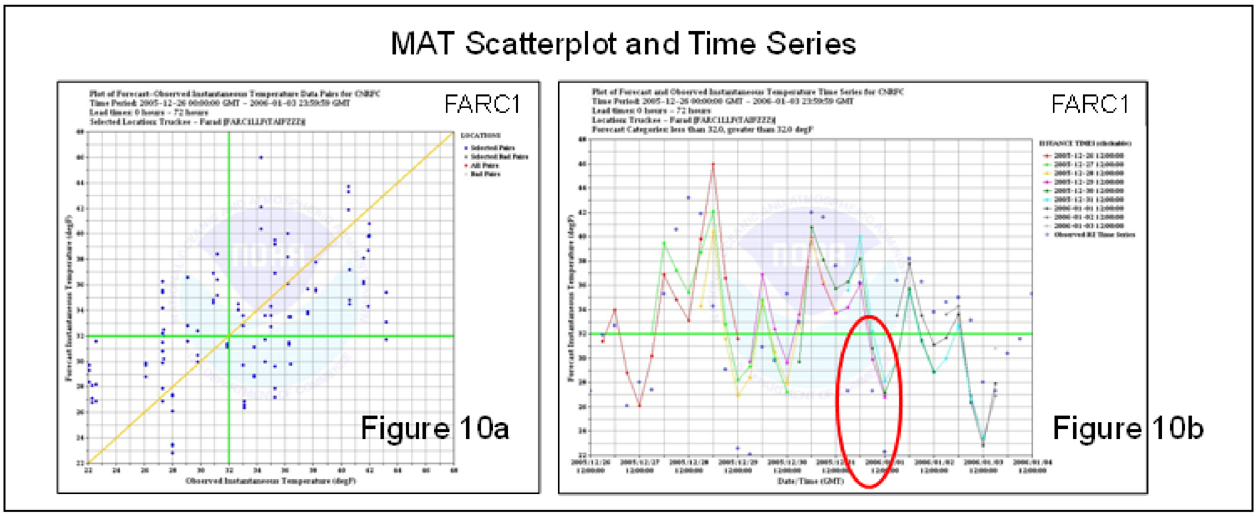
Lead time Statistics – MAP

RMSE appears to increase with increasing lead time for observed category of >= 0.30 inches for HOPC1 for both HAS and HPC (Figures 8a to 8d below; see online Figures 8c and 8d for GUEC1 and Figures 9a and 9b for VISN2 - upper basin). The RMSE is lower for category of <0.30 inches. The HPC plots show fluctuations as lead time increases.

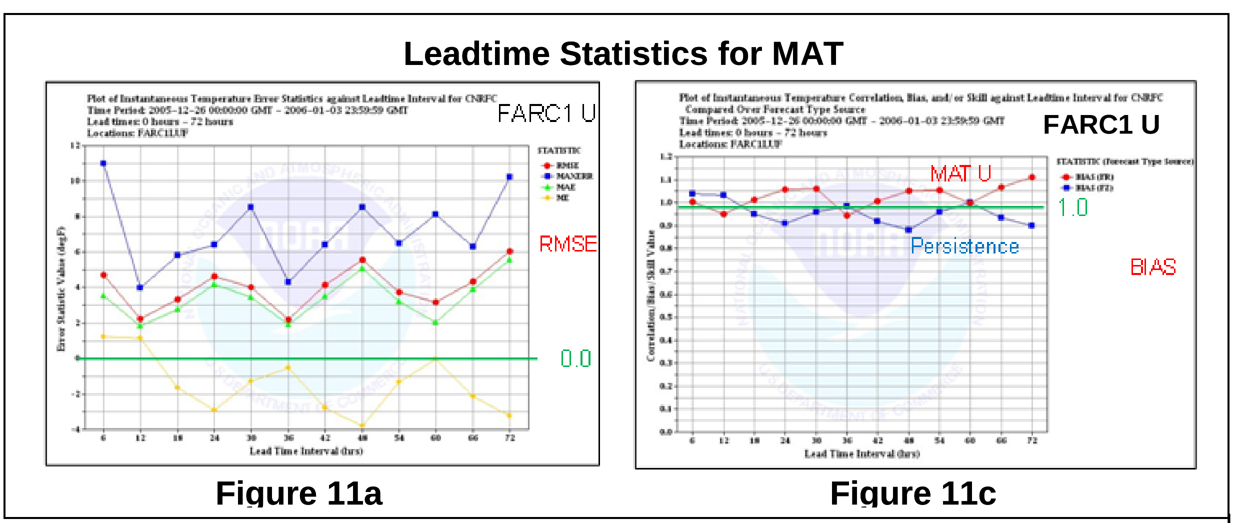


Verification of Forecast Mean Areal Temperature (FMAT): The CNRFC HAS unit produces temperature forecasts using MOS guidance from the GFS model. Forecast MAT were not available for the Russian River drainage for this study. However, temperature may not be a factor in the Russian River basin as all precipitation fell as rain. The following will describe the findings pertaining to the Truckee River FMAT forecasts.

Scatter plots – A scatter plot of observed instantaneous temperature was plotted versus forecast for FARC1 (Figure 10a; see Figure 10b for VISN2 online). There appeared to be no discernable pattern of over-forecasting or under-forecasting.
Time Series – Time series of forecast instantaneous temperature were plotted versus time for FARC1 on the Truckee River (Figure 10b; see Figure 10d for VISN2 online). Plots compared favorably with observed temperature except during 18:00Z 12/31/2005 through 12:00Z 01/01/2006 where temperatures were over-forecast. As can be seen from the scatter plots, the temperatures were above freezing at the downstream point (VISN2); only dropping below freezing only at the upstream FARC1 forecast point. Of particular note is that FARC1 crested just 0.2 feet below flood stage while the downstream points TRRN2 crested 1.5 feet and VISN2 over 5 feet above flood stage.



Plots of Instantaneous Temperature Error Statistics against Lead time Interval were constructed for MAT (Figures 11a and 11b). No discernable pattern is seen for RMSE for both FARC1 Upper and VISN2 Upper as lead time increases. VISN2 U has a slightly higher RMSE at the earlier lead times (6 – 52 hours). No definite advantage of FZ (HAS FMAT forecasts for FARC1 and VISN2 upper basins) over FR (persistence) for BIAS (Figures 11c and 11d) although there is a small consistent over-forecast bias for persistence for the downstream forecast point VISN2 U in Figure 11d.



Categorical statistic (POD) for MAP, HAS versus HPC: the Probability of Detection (perfect score is 1) is best for category 3 (>= 0.50") for Hopland (HOPC1) for both HAS and HPC forecasts (Figures 12a and 12b). POD is best for category 1 (<0.25") for Vista upper basin (VISN2 U) for both HAS and HPC (Figure 13a and 13b), and relatively poor for category 2 (>=0.25" and <0.50") and 3 (>=0.50") for VISN2 Upper for both HAS and HPC. This could indicate the severe under-forecast of QPF for the higher observed amounts at Vista. However, the computed statistics are compromised by the small sample size.

Categorical statistic: Probability of Detection (POD) based on MAP

Category 1: MAP < 0.25"
Category 2: 0.25" <= MAP < 0.50"
Category 3: MAP >= 0.50"

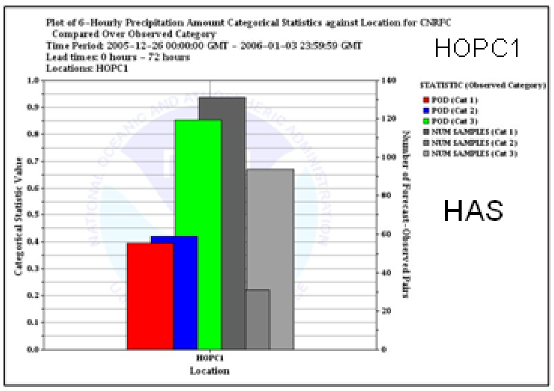


Figure 12a

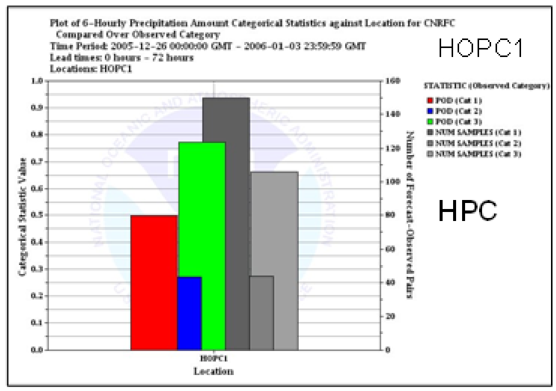


Figure 12b

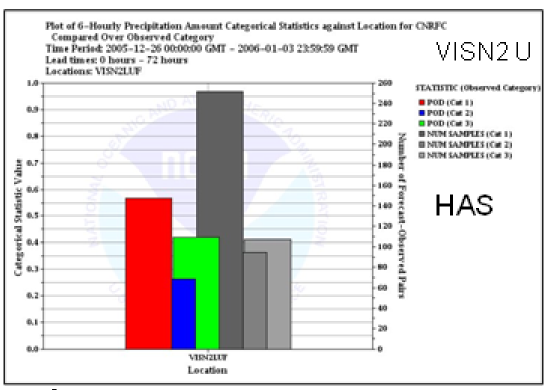


Figure 13a

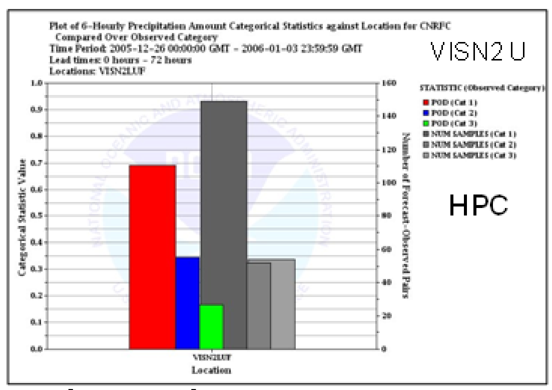


Figure 13b

Bias for MAP based on Type Source (FC: HAS, FW: HPC). There is a tendency for over-forecasting bias for HOPC1, GUEC1 and FARC1 U at the category 1 level (<0.25 inches). VISN2 U has an under-forecasting bias at the category 1 level and a large under-forecasting bias at the category 3 level (>=0.50 inches). No discernable pattern if type source FC-HAS is better than type source FW-HPC. Note the small sample size.

Conclusions

- A limiting factor was the small sample size used in the study. Currently, there is a very small set of historical data available. NWSRFS was not started until 1997 at the CNRFC.

The most complete set of data available at the RFC starts after year 2004. The event in the Truckee basin was characterized by copious amounts of rain falling on the lower elevation basins. Extreme lee-side precipitation events are very rare in the case study area. We would have to go back to January 1-2, 1997 to find an event with similar circumstances for the Truckee. However, orographic-type events are more common on the Russian River basin and similar recent historical events would be easier to obtain.

- The most valuable verification plots generated by IVP are the scatter and time series plots. They show quite conclusively that MAPs were under-forecast during the flood event on the Truckee.

- Based on scatter plot and time series output, the Russian River was generally over-forecast and the Truckee River under-forecast.

- There is a need to verify single flood events. It is true that IVP lends itself well to multiple events of similar characteristics and not single events where small sample size could be a limiting factor. Based on data stored in the CNRFC database, multiple events would be easy to verify for the Russian River, as orographic precipitation events are common in that basin. As just mentioned above, it would be difficult to obtain data for the type of event studied for the Truckee.

- There is an advantage of verifying single storm events of limited duration in that zero QPF is generally a small factor that enter into the statistical computations.

- The CNRFC HAS was slightly better than HPC according to the MAP scatter plots and time series and the lead time statistics for MAP. No decision can be made on who is better according to the categorical statistics for MAP that was computed. Obtaining an adequate sample size appears to be the limiting factor in making a determination using the categorical statistics in this case study.

Additional work to be done on the case study:

- It was recommended that the forecast-observed pairs not be pooled for all lead times since the forecast performance is likely to vary with lead time.

- Expand this case study to include the Napa and Carson Rivers, which also sustained substantial property damage during the 2006 New Year's heavy precipitation event.

- Compare the MAP categorical statistics generated by IVP with the ones generated for the NPVU (National Precipitation Verification Unit) QPF verification.

Unmet needs with current software and hardware

- Need to define a raw model to establish baseline statistics in order to assess value added by the forecaster or hindcasting capability to flush out relative error sources in the model.
- The archive database disk size after the "form-fit-function" upgrade is approximately 900 Gigabytes. I used to run a script provided by the MBRFC called "run_PGbkups" which performed a Postgres backup of the archive tables before the upgrade. I cannot run it now. Perhaps we need to consider purchasing another hard drive of sufficient size as a backup.

References

Herr, H., 2007. Interactive Verification Program User's Manual. NOAA/National Weather Service/Office of Hydrologic Development.

Kozlowski, D. and M. Ekern. December 24, 2005 – January 03, 2006 Southwest Oregon, Northern California and Western Nevada Significant Precipitation Event. NOAA/National Weather Service/California-Nevada River Forecast Center web site, http*://www.cnrfc.noaa.gov.storm_summaries.php*

O'Hara, B., G. E. Barbato, J. W. James, H. A. Angeloff and T. Clyke, 2007. Weather and Climate of the Reno-Carson City-Lake Tahoe Region. Nevada Bureau of Mines and Geology, Special Publication 34. 62, 64.

NWRFC Verification Case Study
Presented on November 18, 2008 by Joe Intermill

Project Summary - Modified from Western Region Hydrologic Verification Team Final Report (June 2008)

The NWRFC examined a rain and snow melt event on the Stehekin River in November 2006. The Interactive Verification Program (IVP) was used with the Archive Database (RAX) to perform statistical analysis on the various forecasts and produce graphics of the results. The peak streamflow forecasts were too low prior to the observed hydrograph peak. In addition to streamflow, the study examined both precipitation and temperature forecasts. Each of these was found to be too low (consistent with low streamflow forecasts). The study also used the Interactive Forecast Program (IFP) along with archived model states to study the sensitivity of the model to changes to temperature and precipitation. By increasing both, the model simulation improved but was still low. The study suggested an examination of the point network used to create basin temperature and precipitation estimates. The study also examined the effects of the actual forecaster modifications made in the forecast period at the time of the forecast. It was determined that modifications made to temperature increased the amount of rain (versus snow) and increased melt. Both of these impacts were beneficial.

The study demonstrated the need for a diagnostic capability in forecast verification. While this study used the IFP, there is a need for a more systematic and objective tool to diagnosis error source. “Time series and scatter plots are (by far) the most useful options within IVP for an individual storm analysis.” “IVP is a very powerful tool that appears to have more utility in a multi-event analysis rather than an individual storm study.” The multi-event or long term verification is an area that needs more exploration.

Case Study Goals

1. The primary goal of the NWRFC National Verification Team Case Study is to assess the validity and usefulness of existing NWS verification and re-analysis tools in the context of a single hydrologic event. Recommendations regarding improvements and the use of these tools will be forwarded to the National RFC Verification Team for review and inclusion in the group report. Requirements for additional tools or the need for expanded capabilities will be included in the report as well. Although the study results could prove useful in the evaluation of many forecast components during the event, this analysis will not be the priority. The aforementioned tools are: Interactive Verification Program (IVP) using the Archive Database (RAX), Interactive Forecast Program (IFP) using archived Operational Forecast System (OFS) FS5FILES, other miscellaneous data viewing and access tools.
2. Develop a plan for systematic testing and implementation of the various verification tools and techniques into “real-time” operations. The hope is that this “real-time” verification information will help to communicate our forecast accuracy in a way that is <u>meaningful</u> and can ultimately be used to improve our forecasts.

Choosing the Study Basin and Event

It was decided that the best way to keep the focus of the study on the verification tools, statistics and methods, was to choose one event at a single river forecast point. The basin/event criteria were:

- Headwater Basin – no upstream routing points or Reservoirs to consider.
- Site should be a routine forecast point with adequate observed and forecast data to assess forcings (areal precipitation, areal temperature) and river forecasts (stage and flow).
- The event should involve river levels exceeding flood stage.
- Model performance and/or forecast quality during the event should be questionable.
- Flooding should be driven by a combination of rain, rain on snow (possible precipitation typing issues), and snowmelt. This will allow a meaningful sensitivity analysis to be performed on the model forcings.

The Study Basin

The Stehekin River at Stehekin, WA was chosen as the study basin. It is a headwater basin located in North Central Washington which drains approximately 321 square miles on the East side of the Cascade Mountain Range. The river finds its way to Lake Chelan just downstream of the forecast point. There is no upstream reservoir or back-water effects to consider. The basin experiences significant snowmelt runoff and can be subjected to heavy rainfall events as well (approximately 80 inches of annual precipitation on average). Much of the basin is forest covered (approx. 80%) with an elevation range from 1099 feet at the gage to 8760 feet at the highest peak. Two to three percent of the basin is covered by a glacier which can have a minor impact by adding to late summer streamflow.

The Event

The study event occurred in November of 2006. A strong tropical moisture feed brought six days of widespread heavy precipitation to basins located on the West side of the Cascade mountain range in Washington and Oregon. Significant rain spilled over the Cascades to impact Eastside basins as well. Ten to twenty inch storm total precipitation reports were common during this event. This rainfall pushed many rivers to record or near record levels.

Temperatures early in the event were near rain/snow thresholds for most of the basin. Precipitation fell as rain at low elevation, while some snow accumulated at the higher elevations. This was short-lived as the tropical air quickly pushed its' way into the Northwest. During the heaviest accumulation periods, the precipitation fell as rain throughout the entire basin.

Although some snow accumulated at higher elevations in the Stehekin basin early in the event, the snow pack was not abnormally large. Representative SNOTEL gages reported less than five inches of snow water equivalent at the time of the heaviest rainfall.

Soil moisture and river flows were normal to below normal leading up to the event.

The Stehekin river gage reported the event peak of 28.8 feet on November 7th, 2006 at approximately 08Z (Midnight Pacific Standard Time). Flood stage is 24 feet and Action Stage is 22 feet.

Methodology

Two separate analyses were performed. The first analysis used the Interactive Verification Program (IVP) while the second employed the Interactive Forecast Program (IFP). Although alternate methods exist (in the form of locally developed applications), tools used in this study were purposely limited to those provided in the nationally supported software suite. The purpose is to identify strengths and weaknesses with these tools and provide that information to the National Verification Team.

IVP Analysis

IVP was designed to link directly to the Archive Database (RAX). Therefore, any analysis using IVP will require that all relevant observations and forecasts reside in the RAX. For the purposes of keeping the study focused, it was decided that the analysis would be performed on Stage, Flow, Mean Areal Precipitation (MAP), and Mean Areal Temperature (MAT).

Datview was the primary tool used to view and quality control data residing in the RAX. Although is has certain strengths, it was too cumbersome to use as a primary tool for data quality control and editing. The inability to view observed and forecast data simultaneously is a glaring weakness as well. Other suggestions for improvements are well documented and will not be listed here. The software was a good “first stab”, but much is needed to make it a viable tool for efficient data viewing and quality control. It should be noted that standard PostGres database queries were used to view data in the RAX as well.

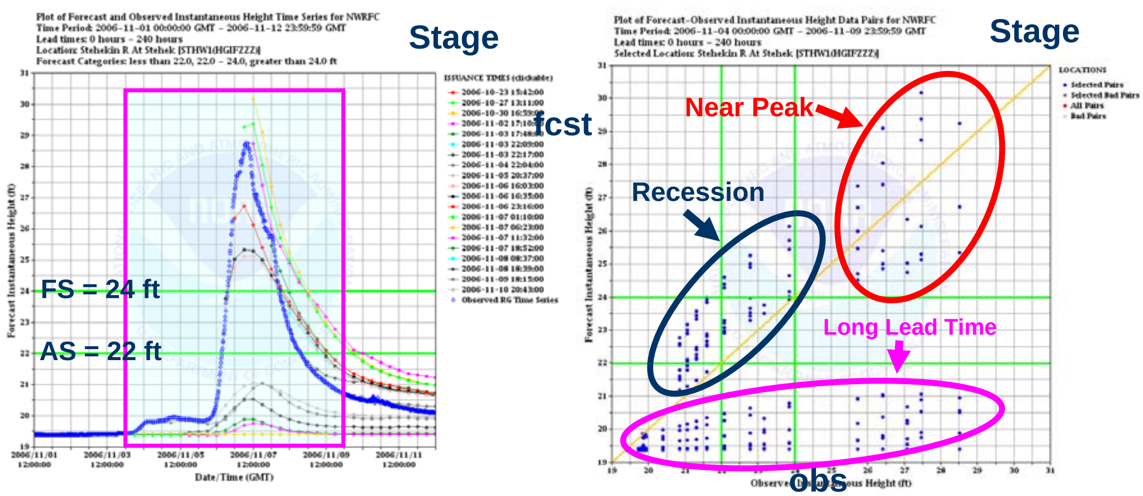
During the review of observed flow data, it became apparent that discontinuities in the time series were the result of rating curve updates. It is not uncommon to receive these updates during event and their impact on verification flow analysis shouldn’t be ignored. Even if rating curves are archived and utilized in the process, these problems will exist. Many areas of the country can access quality controlled streamflow data (USGS) which has these effects “smoothed” and minimized. Some of it is available at very small time steps (hourly or less). Unfortunately, there can be a time delay before this data is available (year or more). The best way to deal with this problem in “near real-time” verification analysis is currently unclear.

The pairing software (ivpruninfo, ivpbatch, and the input file for pairing) was well documented and relatively easy to use. The software is very powerful and provides many options to the user. However, there are some suggested enhancements. Many have been documented by others so I will just mention a couple of the ones that stand out. (1) There is a pre-defined list of SHEF qualifier codes which make observations eligible for pairing. Although this list was developed using good reasoning, there are circumstances that could create a need for using observations with additional qualifier codes. The recommendation is to allow the eligible qualifier codes to be specified in the pairing input file. (2) The pairing window has a lower limit of 1 hour and needs to be smaller if we are to verify quickly changing conditions (i.e. tidal influences).

The IVP is an extremely powerful tool with plentiful options. The standard documentation was very helpful and the examples provided were even more so! Performance issues (speed and memory) really impacted the analysis at times. Certain "workarounds" to access more system memory and temporarily suspend posting to the RAX were somewhat helpful, but overall, the issues remained. Numerous plots were developed to display the various statistical analyses. The flexibility to easily vary selections such as lead time, analysis period and time step is very useful.

<u>IVP Results</u>

Although many setting combinations were exercised, November 4-9 was chosen as the primary Analysis Period. Lead times from 0-240 hours (at a 6 hour time step) were evaluated during the process. For a single event, it was determined that (by far) the most useful plots were the "Time Series Plot" and the "Scatter Plot". See the graphic below for an example of how they can be used together.



- Stage and Flow forecasts were generally low throughout the period. Many issuances near the crest produced high biased recession forecasts.
- The FMAP was very low compared to MAP throughout the period. There were only scattered exceptions.
- FMATs were very bias on the "low side" as well.
- Persistence forecasts were superior to FMATs at times for the short lead times and the very long lead times.
- Forecasts "generally" improved with shorter lead times.
- Statistics/plots are extremely sensitive to the analysis period, lead-time, and number of samples in the comparison.
- Relative error contribution of FMAP, FMAT, antecedent conditions, modeling, forecaster modifications, etc. cannot really be assessed with IVP.
- There is some application for single event analysis within IVP. However, it is probably better suited for multi-event analysis in general.

IFP Analysis

Currently, the best event "re-analysis" tool available at NWRFC is the Interactive Forecast Program provided with NWSRFS. To look at past events, the user must run the IFP on an archived copy of the operational FS5FILES. These files contain all configurations, data, model states, forecaster modifications, etc. that were present at the time of the event in question. For the IFP re-analysis portion of this study, two separate copies of FS5FILES were required:

1. Two days prior to the river crest – this gives us a look at the event in "forecast" mode
   - Analyze soil moisture and snow model states
   - Evaluate the impact of forecaster modifications
2. Two days after the peak – this gives us a look at the event in "observed" mode
   - Look at a "no mods" scenario to evaluate how the raw simulation performed
   - Try to isolate the biggest source of error for the event (soil moisture, snow, FMAT, FMAP)
   - Determine the minimum and most useful mods to simulate the event in an "observed" mode.

IFP Forecast Mode Results

The antecedent conditions were evaluated and found to be reasonable.

- The initial SWE on Nov 1 was:
  - Upper zone: 2.0" [mean elev = 6160']
  - Lower zone: 0.1" [mean elev = 3800']
  - SNOTEL sites in the area had very similar SWE readings on Nov 1.
- Simulated snow "built" to a max of around 6.5" in the upper zone on Nov 7th (once again…this was representative based on area SNOTEL sites).
- Initial simulated soil moisture states were reasonable. Observed flow and SAC-SMA state climatology data were used as references for this assessment.
- Simulated flow was "tracking" the observations and the model states appeared to be adjusted properly.
- There were some forecaster mods which would impact the precipitation typing (i.e. more rain, less snow) during the event. This turned out to be a good move by the forecaster.

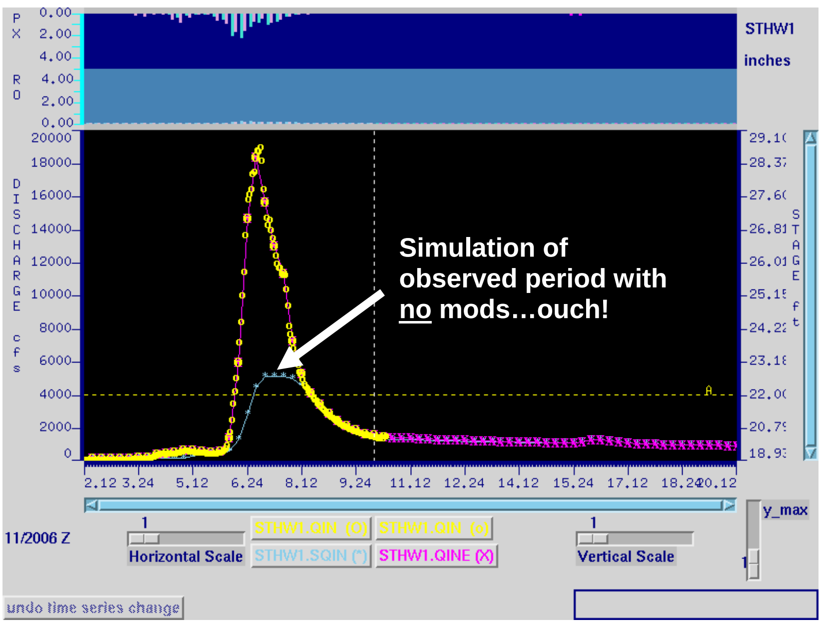
The rise in the forecast looks reasonable given the existing conditions and forecasted forcings.

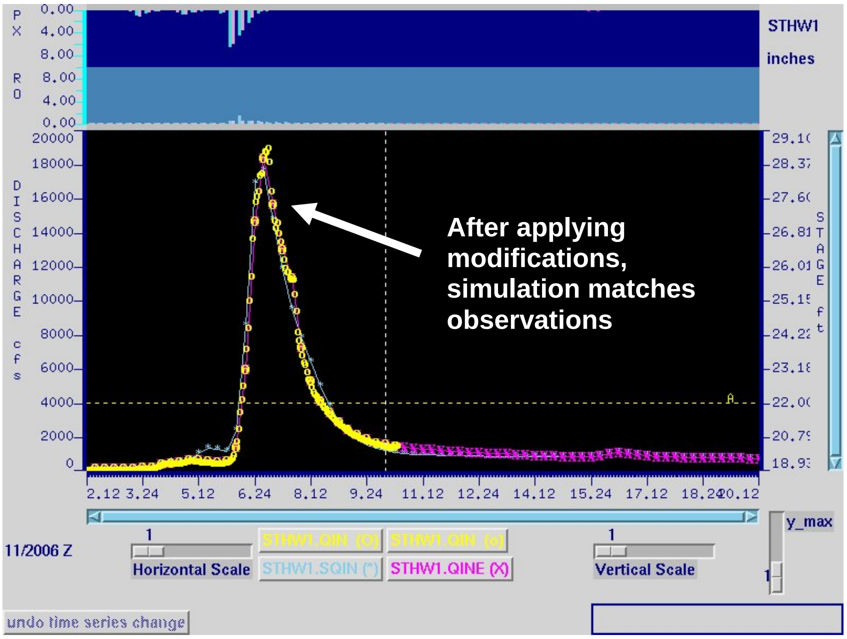
IFP Observed Mode Results

- The "no mods" scenario produced a simulated flow time series that was entirely too low (approximately one third of the observed flow). If we assume that the antecedent conditions were fine (forecast mode analysis), we can reasonable conclude that the combination of "observed" MAP and MAT was totally inadequate. With the potential errors in snowmelt and precipitation typing being relatively low, MAT was ruled out as the "major" source of error.

- To identify the “minimum” error due to underestimated MAP, extreme mods were applied to the simulation for all conditions except precipitation input.
  - Nov 5 : Filled Tension Water “Buckets”
  - Nov 5-8 : Added 20 degrees to all MATs
  - Nov 5-8 : Increased UADJ to maximum (wind effects for rain on snow)

  The resulting simulation was still significantly low. This means that the MAP was low. Possible reasons are the point precipitation input was bad or simply not producing a representative MAP for this event.
- The final task in the re-analysis was to determine the inputs (mods) that would be required to simulate the event. The following was required:
  - Re-distribute the heavy precipitation (13 inches) to concentrate in an 18 hr period starting on Nov 6 at 12Z
  - Add 10 degrees to all MATs from Nov 6 06Z-12Z
  - A unit hydrograph modification was attempted. Although it helped with the rate of rise, the falling limb was still unacceptable so it was not used.
  - Reduce the LZFSC by 20% on the falling limb at Nov 8 12Z.





<u>IFP Re-analysis summary</u>

- Using a very "quick and crude" sensitivity analysis, it is very obvious that forecast mean areal precipitation (FMAP) produced the largest errors in the forecast.
- Low-biased mean areal temperature (FMAT) forecasts had an impact as well. However, the impact was minor compared to the FMAP.
- Antecedent conditions (model states) were reasonable.
  - Simulated snow water equivalent (SWE) values compared favorably with SNOTEL "ground truth".
  - SAC-SMA state values were reasonable based on Fall precipitation, observed flow, and Climatological averages.
- "Observed" MAP and MAT values needed significant modification in the model to simulate the observed flow time series. Point network was not representative.
- A unit hydrograph mod (steeper with less of a "tail") would have helped somewhat, but would only be part of the solution.
- When the rising limb is matched by adding and intensifying precipitation, the falling limb displays symptoms of excess supplemental base flow. This was noticed in some previous events. This should be considered when the basin is recalibrated.
- Although "playing" with archived OFS files using IFP is user friendly and can yield some interesting results, the impacts of point precipitation and temperature data cannot be assessed via this method.
- "Future" temperature Mods produced by a forecaster had a positive impact on the forecast. No other mods had any significant impact on the forecast in a positive or negative way.

Conclusions

- IVP and IFP (re-analysis) have definite strengths and weaknesses.
- IVP is a very powerful tool that appears to have more utility in a multi-event analysis rather than an individual storm study.
- Time series and scatter plots are (by far) the most useful options within IVP for an individual storm analysis.
- IFP is very useful in a post event analysis (requires routine archive of OFS files).
- Using a combination of IVP and IFP can be an effective approach to performing Event Verification. With this combination (or something similar), we can possibly point out inherent problems with our forecasts that could be fixed (i.e. Possibly Improve Forecasts!)

Recommendations

- Exercising the IVP has opened our eyes to many possibilities for verification analysis. This includes current capabilities and possible future enhancements.
- The potential for "real-time" verification using IVP exists. The batch execution mode is being investigated as a method to perform this regular analysis. A pilot project involving QPF is planned at the NWRFC.
- Recent improvements to the Archive system (size, speed) are welcome. However, IVP still encounters performance issues during the analysis of large datasets. Unfortunately, this is where IVP analysis is most relevant. Workarounds have helped some, but the underlying problem needs to be identified and addressed.
- Archived data is critical for many RFC programs (operations, verification, calibration, ensembles, climatologies, etc.).
- As the NWS migrates from NWSRFS to CHPS/FEWS, the Archive must be able to communicate with this new system.
- Effective Data viewing and quality control tools for the Archive are very important "pieces" of a Verification system. We need to continue to improve those tools.
- Migration to the new forecast system (CHPS/FEWS) will require a replacement for the IFP/OFS post-event analysis method. Current capabilities within FEWS could possibly satisfy this need. This possibility needs to be evaluated.

CBRFC Verification Case Study
Presented on March 25, 2008 by Lisa Holts

The Colorado Basin River Forecast Center case study stemmed from the Western Region Verification Team Study. The study focused on the event on February 10-13, 2005, which produced widespread precipitation over Arizona. The event persisted three days and caused rain amounts to reach 3 inches in some areas. The Gila basin in east central Arizona and Verde basin in central Arizona were examined with two headwaters and two downstream points in each basin. Five of the segments examined reached flood stage. The segments were calibrated on a one hour time step with MAPX input from MPE.

The NWSRFS OFS was run in a re-forecast mode to create raw model forecast times series and simulated time series. It was started from a set of fs5files saved February 7, 2005 and run forward through February 18, 2005 at six hour time steps. No mods were included during the run period. The raw model forecast time series was created by feeding 72 hours of QPF, along with the observed temperature and freezing levels into the future. The raw model simulated time series was created by simply feeding observed precipitation, temperature and freezing levels. Therefore an easy comparison could be made that looked strictly at the effects of the 72-hour QPF values on the model.

The MAPs show that the Gila basin precipitation was mainly under forecast especially with amounts greater than 0.15 inches. In the Verde basin, the errors in the MAP forecast had a much wider spread with amounts less than 0.6 inches having nice scatter, but amounts greater than 0.6 inches under forecast. (Figures 1, 2)

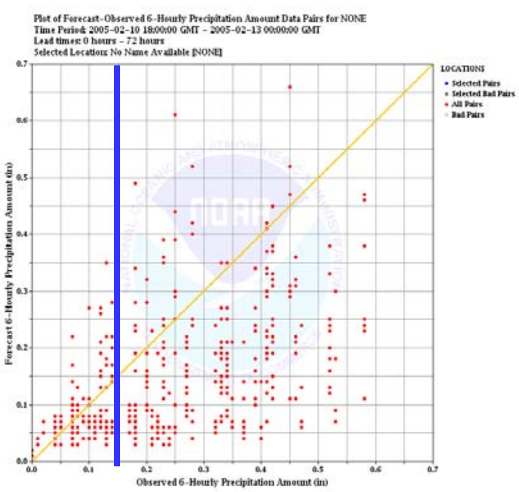


Figure 1. Gila basin MAP indicating under forecast.

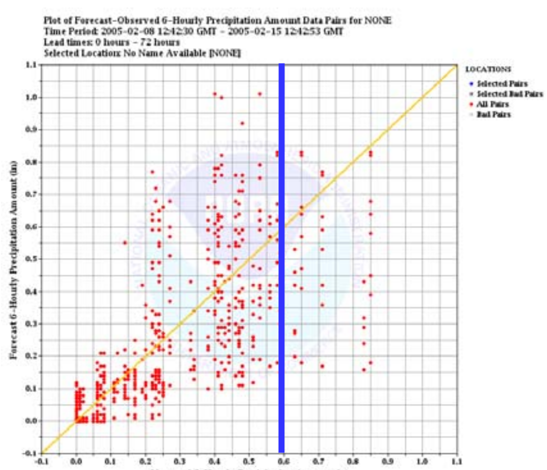


Figure 2. Verde basin MAP indicating nice spread with slight under forecasting.

The error seen in the MAP data was then translated to the streamflow forecasts. In the Gila basin, the under forecast MAPs were reflected in an under forecast in streamflow, while in the Verde the better spread created slightly better streamflow forecasts, but are still slightly under forecast.

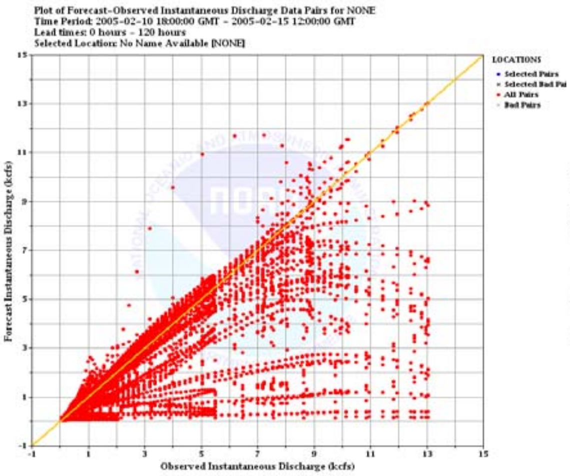


Figure 3. Gila streamflow forecast versus observed; under forecast.

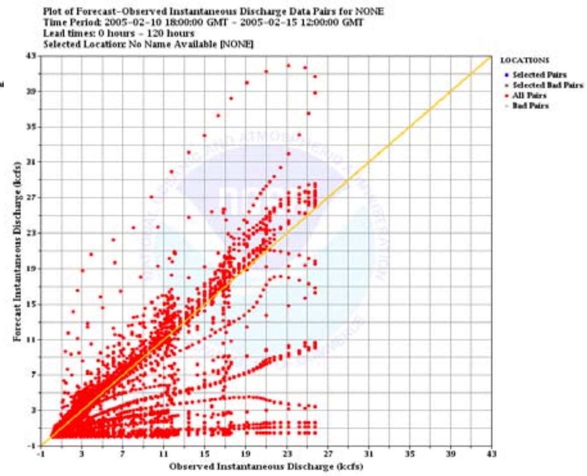


Figure 4. Verde streamflow forecast versus observed; slight under

In the Gila basin (figure 5), each segment was under forecast. The pink line depicts the actual observed streamflow through the segment while the blue dotted line depicts the simulated time series from the raw model simulation. The large difference between the blue and pink lines is likely representing not enough precipitation getting into the model from the previous events. Also, there are a lot of gaps in the radar and gage data over the Gila basin that likely missed the precipitation that occurred previously as well as during the current event. The Verde basin (figure 6), has the opposite effect. The raw model simulations as well as the raw model forecasts are over forecast. The difference in the Verde is that there is very good radar and gage coverage allowing more precipitation to get into the model. It is also possible that too much precipitation was getting into the model in this area leading up to and during the current event.

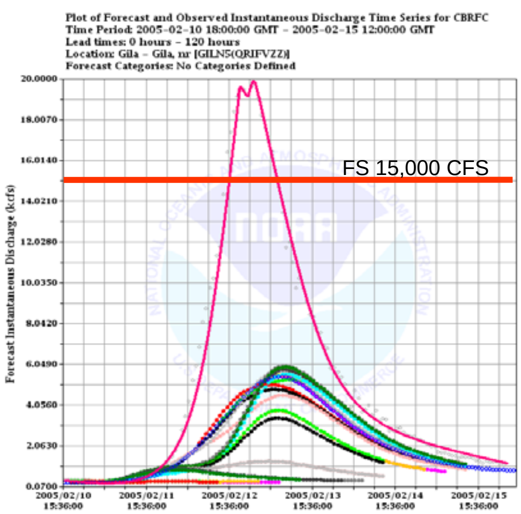


Figure 5. Gila streamflow forecast under forecast.

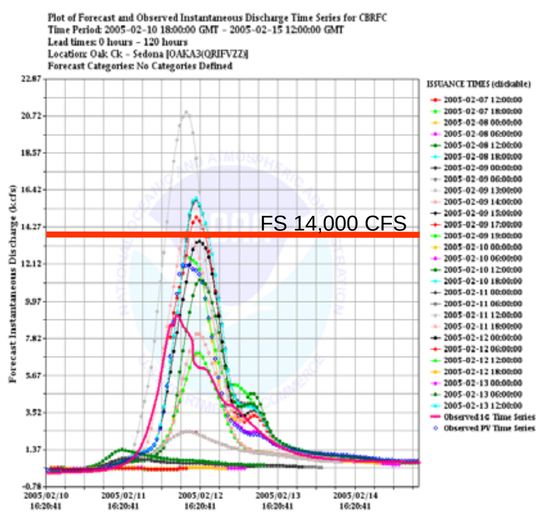


Figure 6. Verde streamflow forecast over forecast.

The statistics in this study were all very dependent on the observed mean at each segment. The larger the mean led to larger statistical errors as illustrated in figure 7 with the

Verde basin. The errors fluctuated with the observed mean and also increased slightly with lead time. In figure 7, there is an increase in the short lead time of about 18 hours. This corresponds nicely with an increased error of the MAPs and thus showing that the error found in the MAPs is present in the streamflow errors as well.

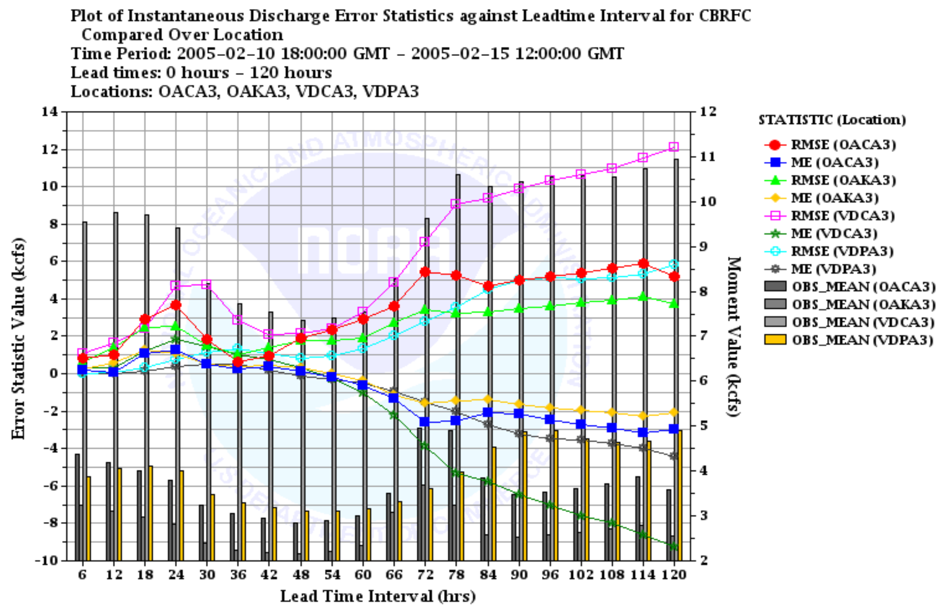


Figure 7. Verde streamflow RMSE and ME plotted with the observed mean where you can see the dependence.

It was also discovered that the errors found in the entire set of data was very dependent on the rising limb of the hydrograph. Figure 8 shows the bias over the entire set of data while figure 9 has separated out just the rising limb. The rising limb was determined by when the hydrograph began to rise to when it initially peaked. The fluctuations seen in the entire set are from the fluctuations seen during the rising limb.

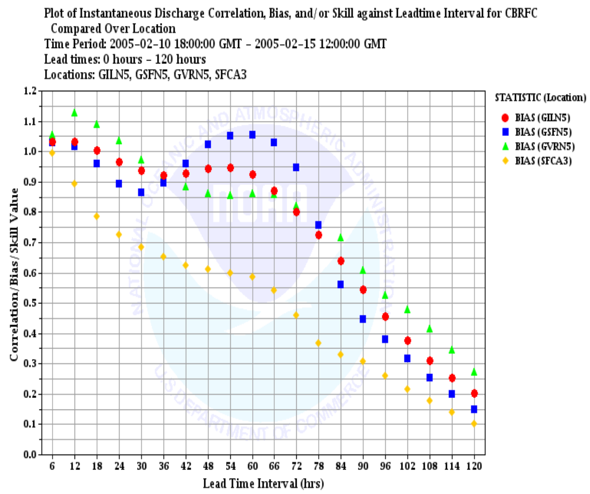


Figure 8. Gila streamflow bias over the entire forecast period.

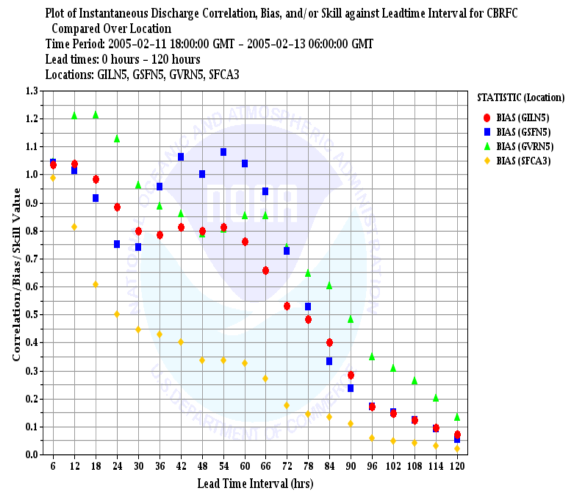


Figure 9. Gila streamflow bias over the rising limb of the hydrograph.

The CBRFC case study found that the flow forecasts were highly dependent on the FMAP but better QPF in the Verde Basin than in the Gila Basin did not translate into better flow forecasts. The incorrect QPF does not fully explain the poor flow forecasts in either basin. It points to needing to address the observed precipitation and possible MPE issues as well as previous model states especially focusing on soil moisture.

To better understand the flow forecasts, it is best to break out the rising limb and focus on those sources of error since they are causing the most error over the full period. Also, when comparing error, it is important to keep in mind the observed mean flow to better interpret the results between segments as true differences or possible differences from the flow magnitude.

LMRFC Verification Case Study
Presented on October 28, 2008 by Kai Roth

Synopsis

Hurricane Katrina was an Atlantic formed hurricane during the 2005 season. It was the costliest as well as one of the five deadliest storms in the United State's history. Katrina formed over the Bahamas on August 23, 2005 and began moving west over the Florida peninsula as a category 1 storm on August 25th. It continued its westward movement over the warm waters of the Gulf of Mexico. A ridge over Texas caused the storm to turn more northward. It quickly gained strength to a category 5 storm on the morning of August 28th with winds of over 170 mph. In general, a category 5 storm cannot maintain their intensity for long. So was the case for Katrina. The storm diminished in strength to a strong category 3 storm before it made landfall on August 29th south of Buras, LA. Wind speeds at landfall were 125 mph with a central pressure of 920 mb recorded at Grand Isle, LA (3rd lowest on record for an Atlantic land falling hurricane).

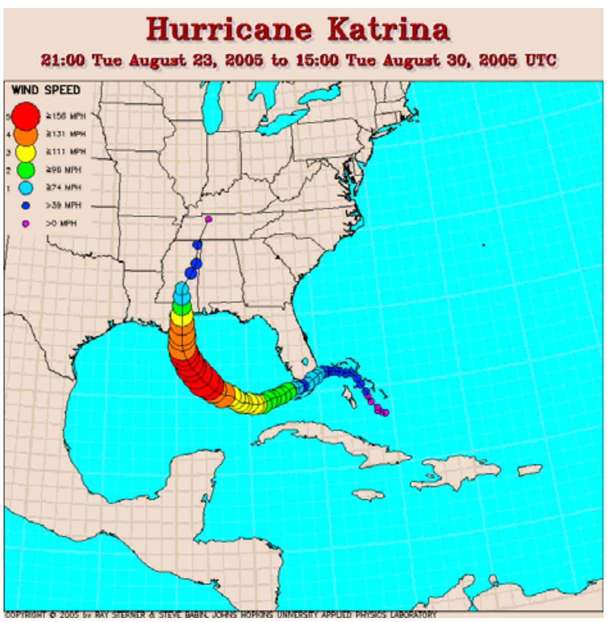


Figure 1: Path of Hurricane Katrina with intensities.

Rainfall Totals

As Katrina approached land, rain bands began affecting the gulf coast. Radar estimated rainfall accumulations exceeded 8 – 10 inches in areas (fig. 2). As the storm moved northward, the rainfall totals diminished, but rainfall amounts still exceeded 2 – 4 inches through the Mississippi and Ohio Valleys.

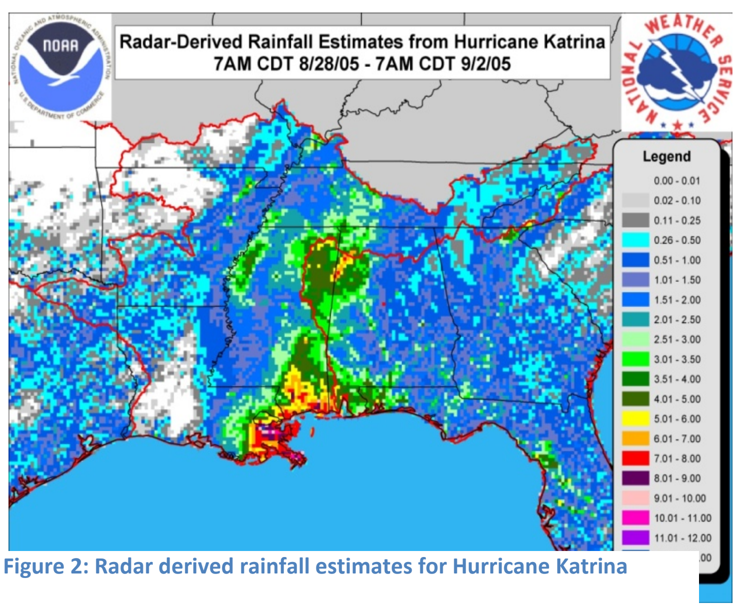


Figure 2: Radar derived rainfall estimates for Hurricane Katrina

Storm Surge

While Hurricane Katrina made landfall as a category 3 storm, the storm surge associated with this storm was more representative of a category 5 storm. Because of the size of the storm and the rapid decrease in intensity, the storm surge did not have time to dissipate. Upwards of a 30 foot storm surge occurred along the Mississippi Gulf Coast while a 15 – 20 foot storm surge was observed in New

Orleans (figure 3). This surge presented a unique problem for rivers in the area. Instead of a flood wave flowing down stream, it flowed upstream.

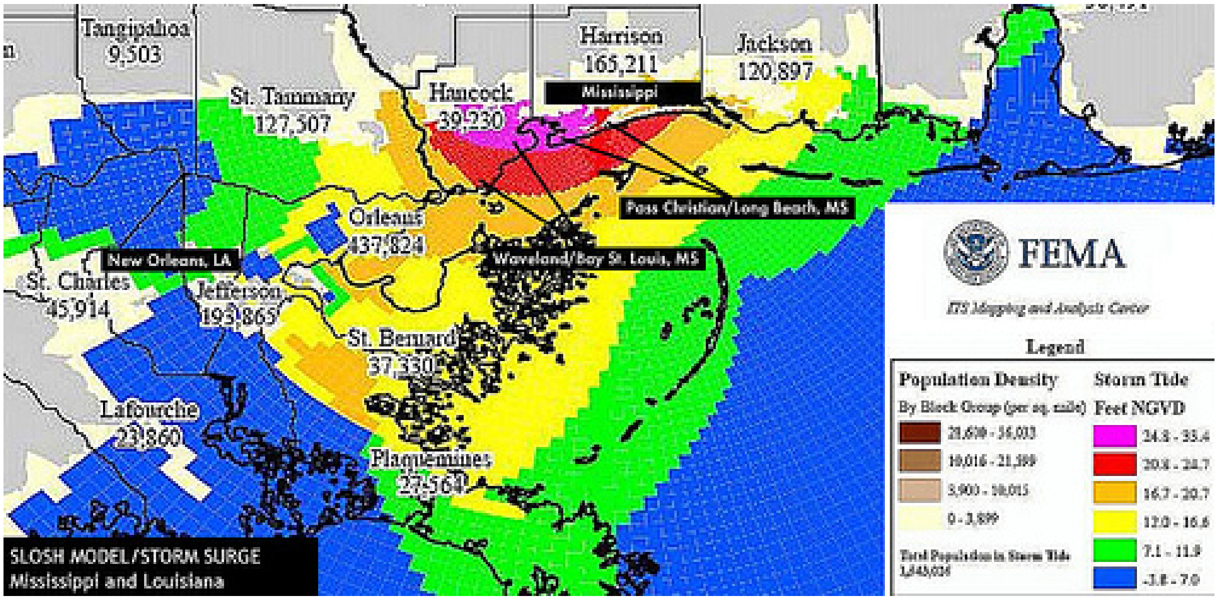


Figure 3: Storm surge associated with Hurricane Katrina

Another challenge in verifying river forecasts is that most coastal river gages were destroyed by the surge. The key to verifying river forecasts was to find gages that remained working throughout the storm. The only ones that remained reporting along the gulf coast and were greatly impacted by the storm surge were the gages near the mouth of the Mississippi River (fig 4).

Figure 4: Forecast points to the verified.

River Verification

The Mississippi River is a slow moving, slow responding river in general. Under normal circumstances the crest at New Orleans will be known 10 – 14 days in advance. When a hurricane is in the forecast, we may only get 24 hours of lead time at points near the mouth of the Mississippi River. This is because the storm surge at these points can rise and fall in 24 hours or less.

The four points depicted in figure 4 were verified against the forecasts on the Mississippi River for the period 2004 – 2008. Because of the challenges of dealing with storm surge, statistics were generally higher for Hurricane Katrina than the 2004 – 2008 period. RMSE were 1-2 ft. higher for the hurricane than the comparison period. MAE were upwards of 1 ft. higher for Katrina, while ME was approximately .5 ft higher for Katrina. Historically, correlation approached “1” for all points studied, which is the desired score. Correlation for Katrina was poorer the further downstream you went. This is due to the storm surge moving and attenuating upstream. Similarly the Bias was large for Katrina as you moved downstream towards New Orleans. This again was due to the storm surge. Persistence performed better than the forecasts for Katrina. This was also due to the influence of storm surge.

Conclusions

While statistically forecasts for Hurricane Katrina were not as good as the cumulative statistics from 2004 – 2008, the forecasts were acceptable given the extenuating circumstances.

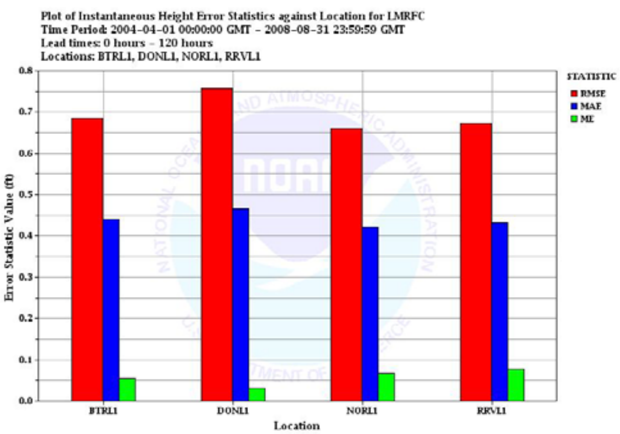


Figure 5: RMSE, MAE, ME from 2004 - 2008.

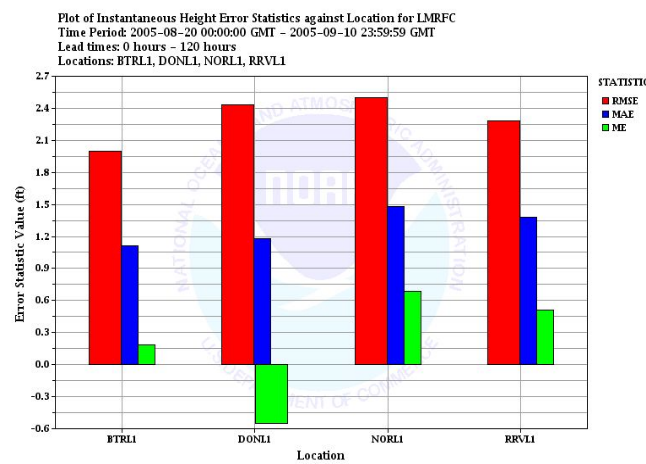


Figure 6: RMSE, MAE, ME from Hurricane Katrina.

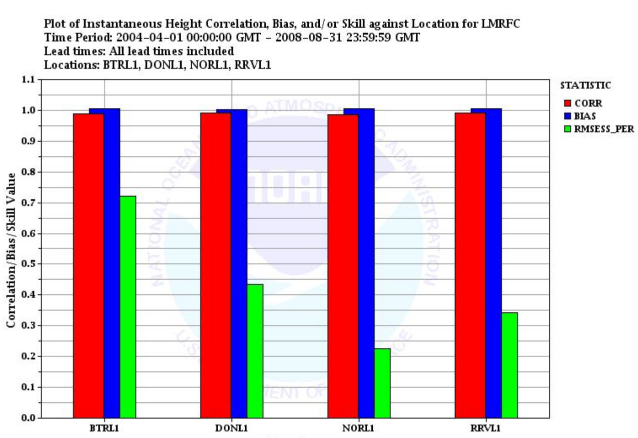


Figure 7: Correlation, Bias, and Skill from 2004 - 2008

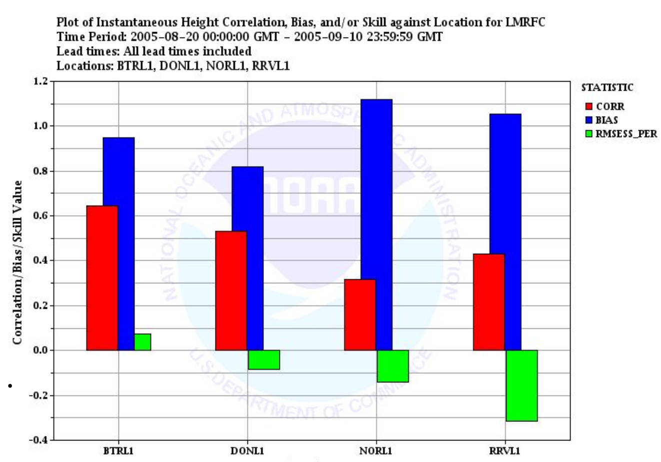


Figure 8: Correlation, Bias, and Skill from Hurricane Katrina

NCRFC Verification Case Study
Presented on November 18, 2008 by Mike DeWeese

Background

2008 was an extremely active flood year for the NCRFC, with major flood events starting in January and continuing through the disastrous record flooding over the Midwest in June. In total, flooding activity so far this year has produced in excess of 60 new records in our area or responsibility (figure 1). The heavy operational demand on staffing resources resulted in delayed projects as a consequence, including verification. However, beginning in late September we were able to initiate verification activities using the Interactive Verification Program (IVP) to analyze the preponderance of forecast data generated this year.
Our initial goal was to establish a baseline set of reference metrics for RFC forecaster and WFO reference. However, following the unusually large number of record events we realized we had a valuable opportunity to investigate how well our forecasts verified under extreme flow conditions. The final goal was to gage the performance of the NCRFC QPF contingency ensembles based on experimental HPC max/min QPF guidance.

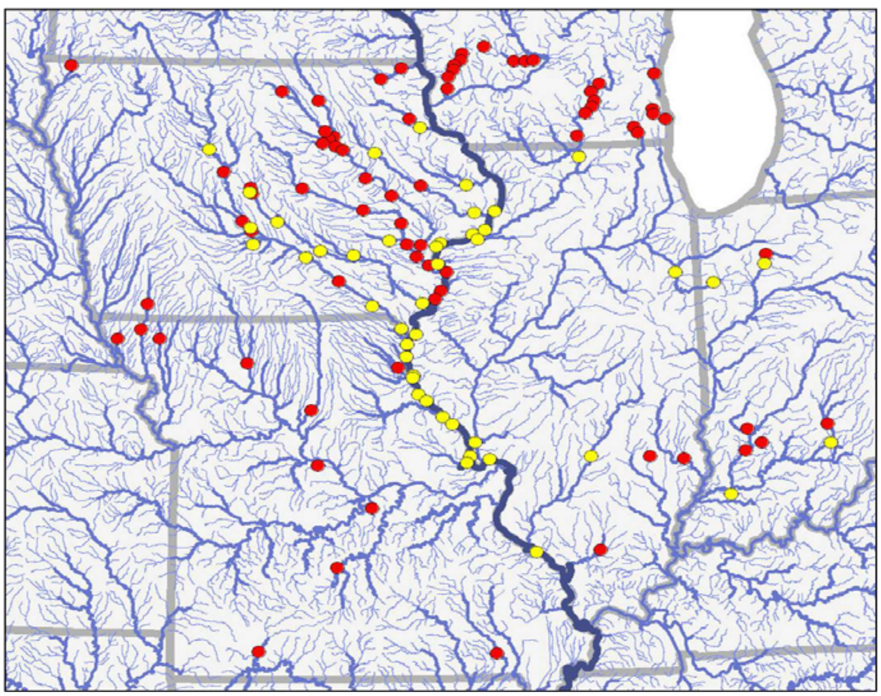


Figure 1: Record and near record floods from the summer of 2008.

Methods

To simplify the analysis, we chose four locations on the Cedar and Iowa Rivers; Waterloo (ALOI4), Cedar Rapids (CIDI4), Iowa City (IOWI4) and Columbus Junction (CJTI4). CJTI4 is below the confluence of the two rivers and was chosen to determine if there a cumulative error effect at downstream locations. All four sites had record flooding this year. We used two

metrics, Mean Absolute Error (MAE) and Mean Error (ME), as a baseline measure of forecast quality.
We conducted three different analyses to meet the project goals:

1. Analysis over the two year archive period to measure our overall performance.
2. An analysis of the record flood event in June 2008.
3. An analysis of the QPF contingency ensembles during June 2008.

Results

In all locations the error statistics show a similar trend over the archived period of record. Forecast errors are generally low from days 1-3, and then increase substantially on days 4-7 (figure 2). The ME also indicates relatively little bias during the first half of the forecast window, with an increasing negative bias beyond day 3. These general trends are consistent with results from the record flooding in June of 2008. We also saw the same trend when analyzing timing effects, i.e. before the crest vs. after the crest. Although the magnitude of error prior to the crest is two to three times greater than after the crest.

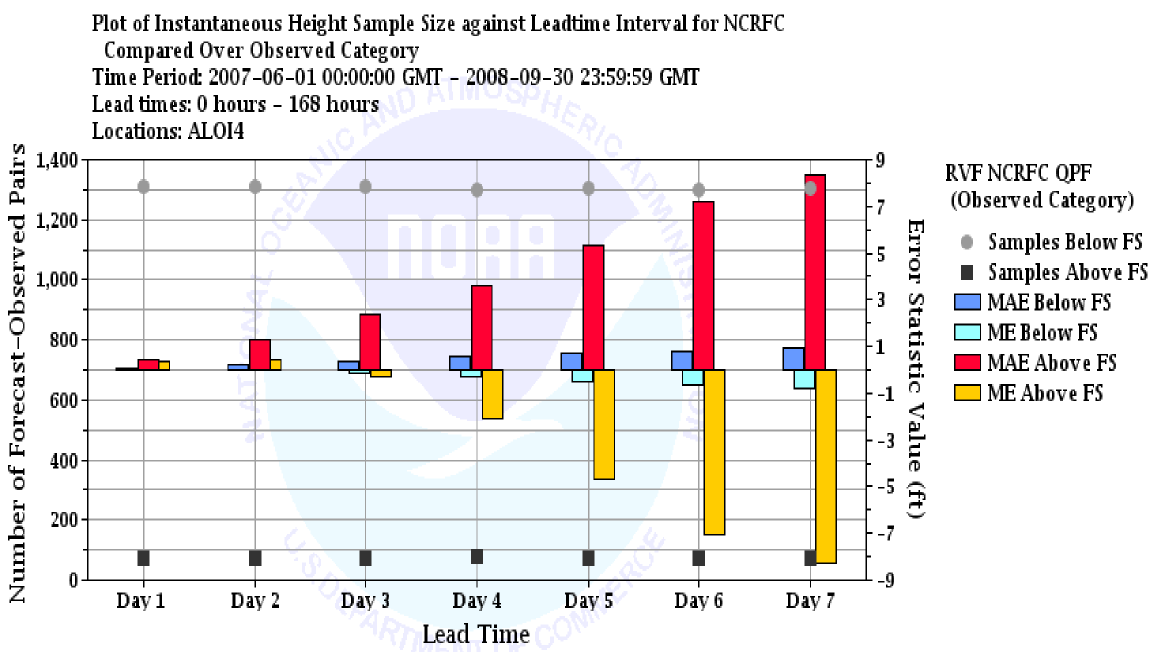


Figure 2: Error statistics above and below flood stage at Waterloo, Iowa.

The same analysis was conducted on the NCRFC QPF ensemble hydrographs for the record flood event in June, 2008. Error statistics were calculated for ensembles using 24, 48, and 72 hour forecast QPF and 24, 48, and 60 hour durations of maximum and minimum QPF at the 95% confidence interval. The results are more difficult to interpret when plotted because of limitations identified in the IVP related to the interpretation of SHEF type source codes (figure 3). However upon close inspection of these results we concluded the following:

1. There is less error using 24 hour forecast QPF than using no QPF.
2. Errors decreased using longer durations of forecast QPF.
3. At major flood levels there was less error from the maximum QPF ensembles than the official forecast QPF ensembles.
4. The minimum QPF ensemble error is constant regardless of the QPF duration used.

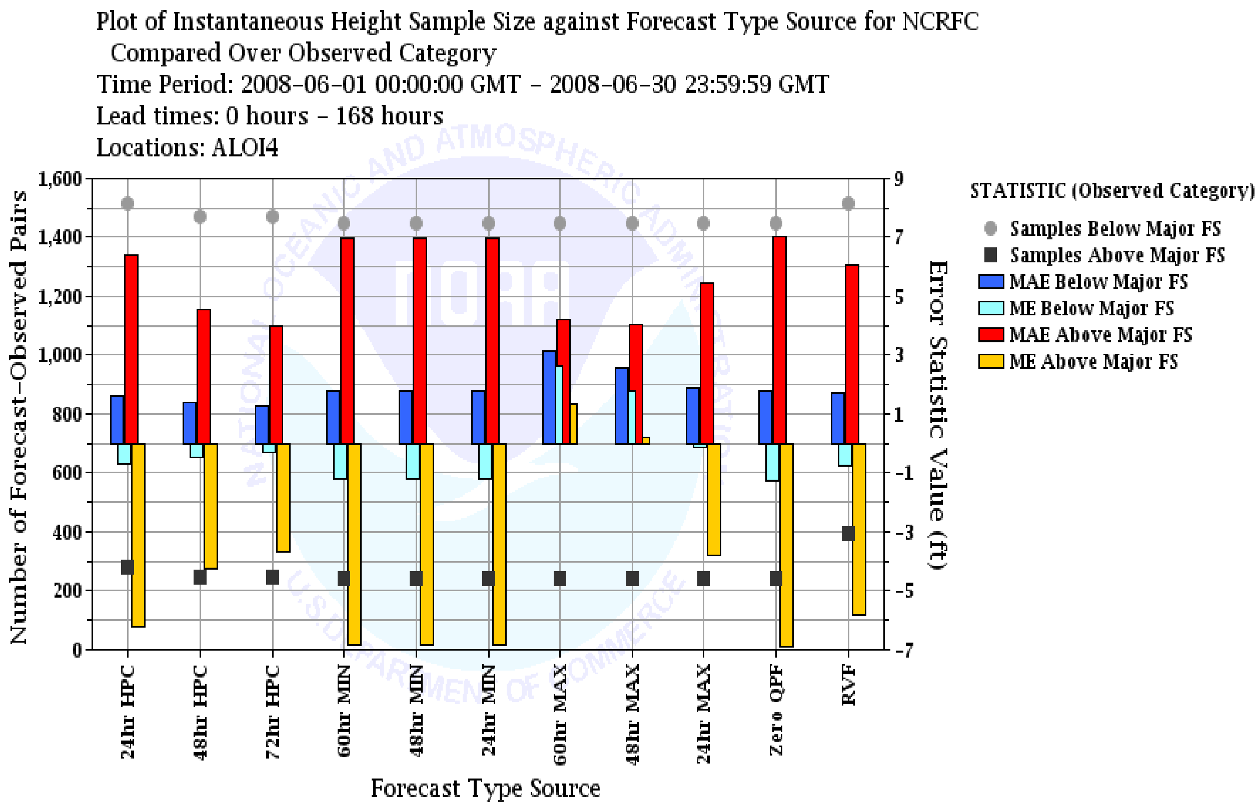


Figure 3: June, 2008 NCRFC QPF Contingency ensemble error by qpf type and duration above and below major flood levels at Waterloo, Iowa.

Conclusions and Future NCRFC Activities

The increasingly low bias in the forecast window suggests we need to use longer durations to support a 7 day forecast. However, as bias is reduced we also see indications of larger errors. Further analysis is planned at the NCRFC to determine if adding longer duration QPF will provide more forecast accuracy in larger sample sets.

We unexpectedly saw no significant increase in error using the maximum QPF scenarios at the 95% confidence interval. This may be due to the magnitude of the June flood event. Further analysis will be conducted on the QPF ensembles to determine if this trend is evident over the entire period of record, and in other basins.

Looking at the results from upstream to downstream points, we found no cumulative error effect. In fact, the error appears to decrease at downstream points.

We found using the IVP in batch mode to be more useful when conducting analysis of more than one or two basins. This is especially true when generating routine statistics. However, due to the number of possible variables in the input files it is very easy to miss one required variable change and get the wrong results. We plan to develop a local application to provide an operational IVP batch menu to reduce potential editing errors and improve operational efficiency.

The NCRFC AOP for FY09 includes providing real time verification on a public web page for the Meramec River. This will be used as a prototype for future plans to provide real time verification to the user community for all basins.

The IVP has numerous options and capabilities we found useful, although some limitations remain especially in generating meaningful graphics. For example, reference figure 3. There currently is no user control to arrange the x axis ordinates in a consistent way.

An original attempt at probabilistic verification using a beta release of the EVS program showed limited success due to a lack of sufficient sample size. (Several years of probabilistic forecast data was lost when the NCRFC archive server failed.) There is a large demand for this information in our user community. Therefore, we plan on initiating another effort at probabilistic verification this year by pooling samples across a regional scale.

MBRFC Verification Case Study
Presented on October 25, 2008 by Julie Meyer

1. Introduction

1.1 Background

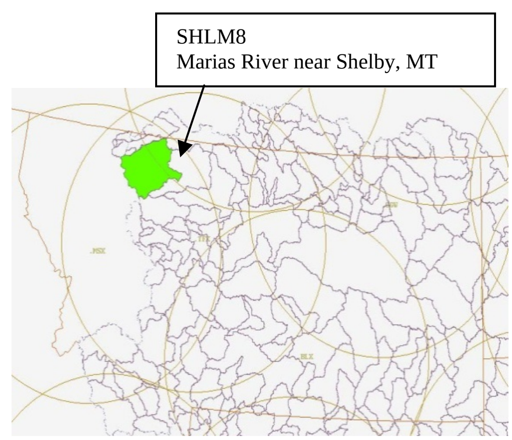
Several years ago the weather forecast office in Great Falls, MT (WFO-TFX) asked "How good are the daily river forecasts?" That question was addressed by utilizing a local verification application which had been used for several years for year-round daily forecast points. This application produced monthly statistics including mean absolute error (MAE), mean error (ME), variance, and standard deviation. Six stations were examined in this earlier study, these were: Gallatin Gateway, MT (GLGM8), Shelby, MT (SHLM8), and Toston, MT (TOSM8) from the Upper Missouri forecast group and Corwin Springs, MT (CORM8), Livingston, MT (LIVM8), and Billings, MT (BILM8) from the Yellowstone forecast group. These locations are all forecast seasonally from April thru June. The local application produced monthly MAE for each station for the years 1998 thru 2001. Statistics were looked at by station, by forecast group, and using all 6 stations combined. The results of this study were inconclusive.

1.2 Current Case Study

The current study re-examines WFO-TFX's question "How good are the daily river forecasts?" using the new national verification software IVP. Four locations were considered for examination, including three in the upper Missouri forecast group and one in the Yellowstone forecast group. From these, Shelby, Montana (NWS ID "SHLM8") was selected for detailed evaluation.

The Shelby river gage is located on the eastern side of the Rocky Mountains along the Hi-Line in northern Montana. Drainage area is 3.243 square miles with approximately 518 square miles of noncontributing area. Currently it is modeled with two elevation zones, these are:
SHLM8UPR – the 5.8% of drainage area above 6500 ft to 8832 ft MSL

SHLM8LWR – the remaining drainage area between3087 ft to 6500 ft MSL



Flows are regulated to a degree by Lower Two Medicine Lake, Four Horns Reservoir, Swift Reservoir and Lake Frances. Irrigation diversions supply about 50,000 acres upstream and 15,000 acres downstream of station.

2. Procedure

2.1 Data Used

Data for April, May, and June of each year during the period 1998 thru 2008 were used for the broader analysis of stages (observed data and forecasts issued), In addition, for the years 2004 thru 2006 both flows and stages were analyzed. Computed observed hourly flows were obtained from the Montana USGS for the flow analysis.

It should be noted that during the study period, Shelby MT only went above flood stage twice, June 2002 and May 2008. The State of Montana in general had below normal snow packs for the spring snowmelt season during the analysis years.

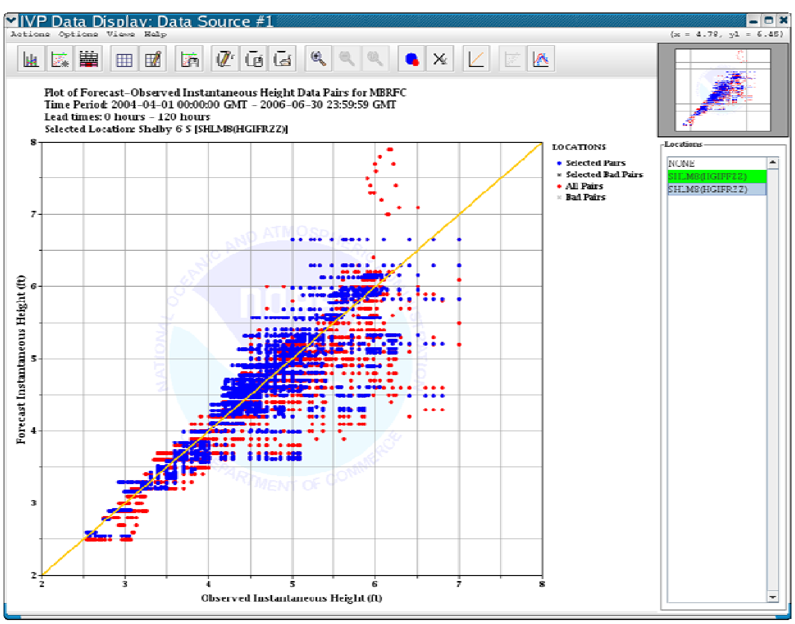
2.2 The Stats

This case study was used to explore the new features and statistical tools of the IVP application, and to gain knowledge on how to interpret these statistics.

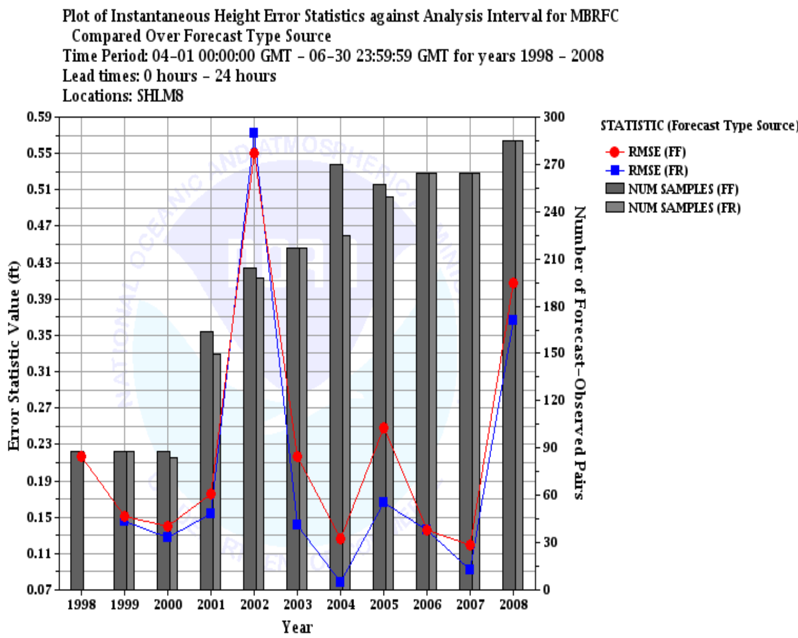
Using the IVP application the following statistics were generated:

- Scatter plots
- Sample size
- RMSE
- MAE
- ME
- Pearson's Correlation
- Bias
- RMSE-SS

3. Results



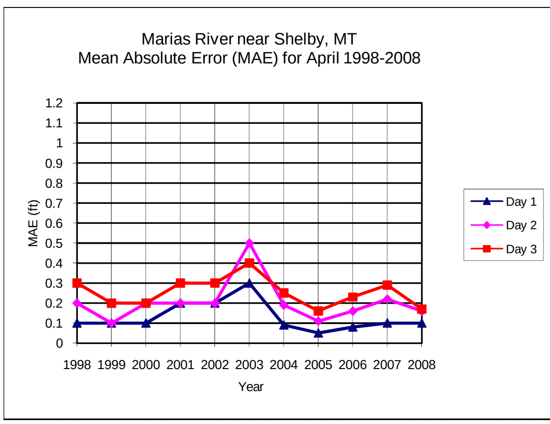
This case study compared the River Forecast Center (RFC) forecasts (FF) with persistence (FR). The scatter plot indicates reasonable positive correlation.

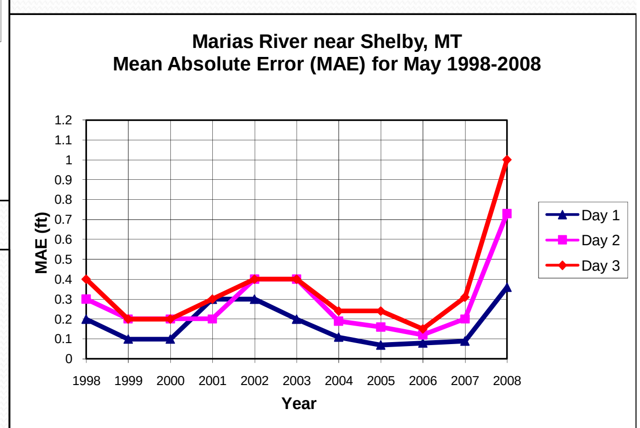


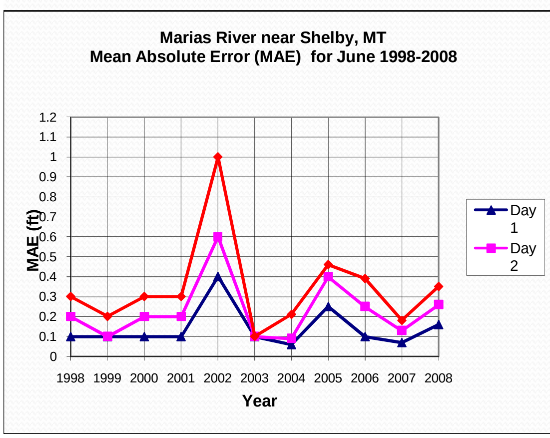
The RMSE for RFC forecasts (FF) is in red and the persistence (FR) is in blue. The gray bars indicate sample size.

This plot indicates that for the past ten years persistence is slightly better than the RFC forecasts.

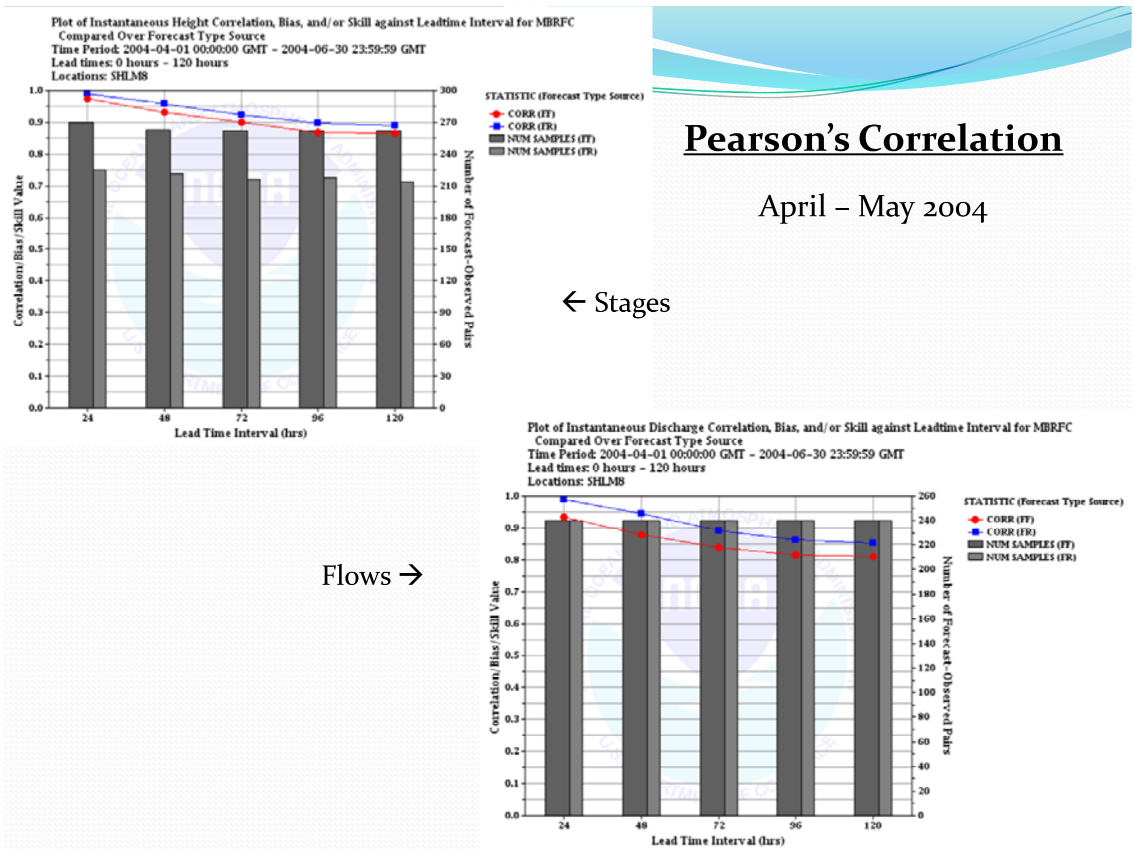
MAE was also examined by month for each year (charts below). Again, persistence proved to be slightly better than RFC forecasts.





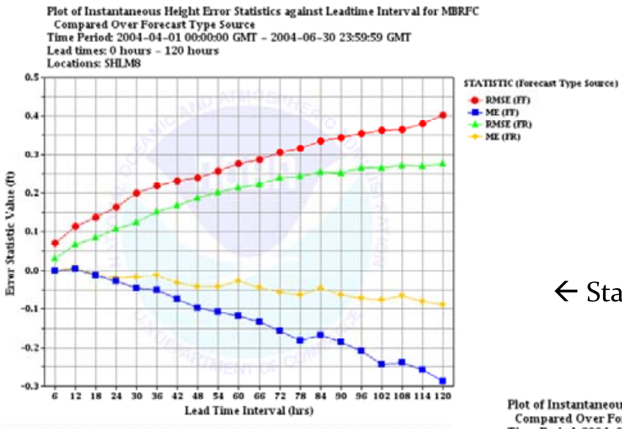


Both RFC flow and RFC stage forecasts were compared to persistence for the years 2004 thru 2006. Plots for 2004 are shown with one exception RMSE-SS for stage forecasts in 2006 is also included. Plots and results are similar for 2005 and 2006.



The plots for Pearson's Correlation (see above) indicate that persistence is again slightly better than the RFC forecasts for both stages and flows. The plots also indicate decreasing correlation with increasing lead time.

The plots for the RMSE and ME (see below) indicate that the RMSE error increases as one goes further out in time and the forecasts tend to under forecast (negative ME), which gets worse with increasing lead time.



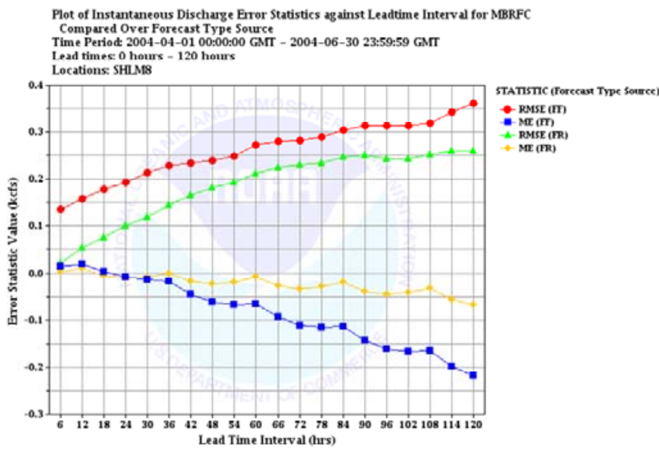
Root Mean Squared Error (RMSE) and Mean Error (ME)

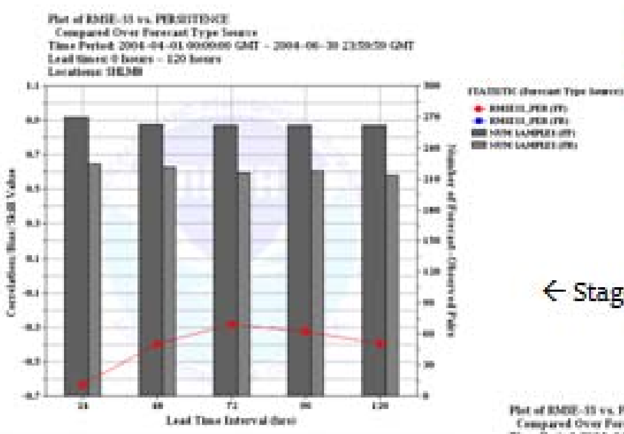
red & blue = FF
green & yellow = FR

← Stages

Flows →

April-May 2004



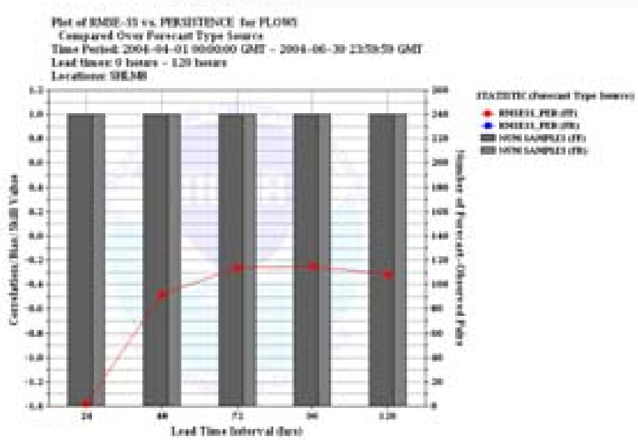


Root Mean Square Error Skill Score (RMSE-SS)

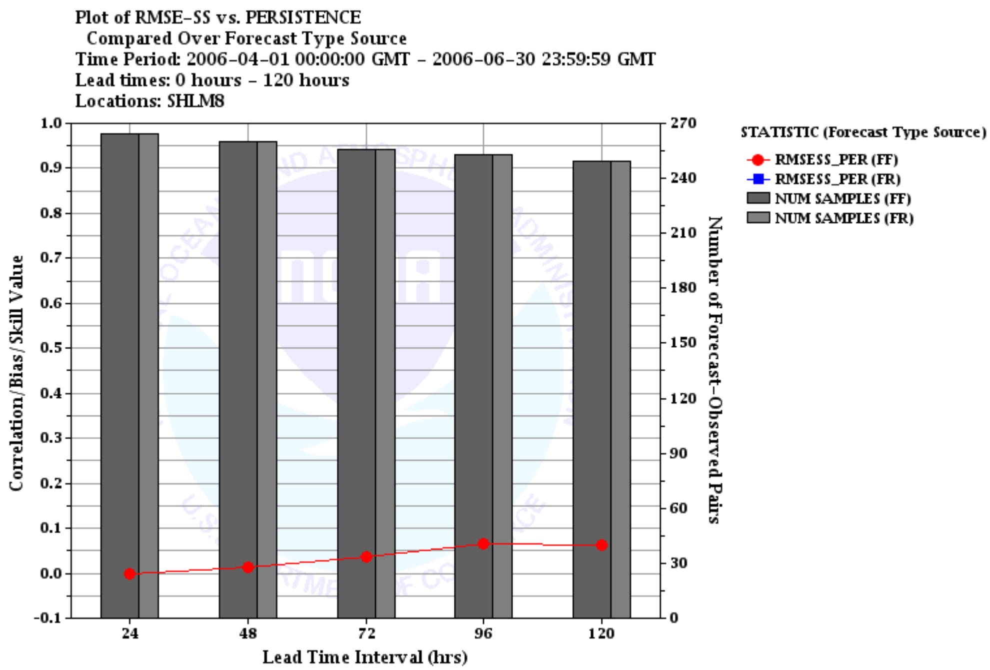
← Stages

Flows →

April-May 2004



Skill scores (see above) are less than zero for both flow and stage forecasts which indicate RFC forecasts are less desirable than the reference, in this case, persistence.
However, for the 2006 stage forecasts the RMSE-SS (see below) indicates the RFC forecasts were slightly more desirable than persistence.



4. Conclusions

Currently it appears that persistence forecasts are for the most part slightly better than the RFC issued forecasts. This is mostly likely due to below normal spring snowmelt seasons for most of the study period. This case study re-enforces the need to break up this basin and take advantage of the USGS/USBR gages, break out the reservoirs and take into account off-stream storage and diversions.

This case study produced more questions than answers, so further analyses at more forecast points will be performed. Forecast points that are further downstream will be emphasized. In addition, as time permits, will expand the Shelby, MT to include other statistical measures such as categorical.

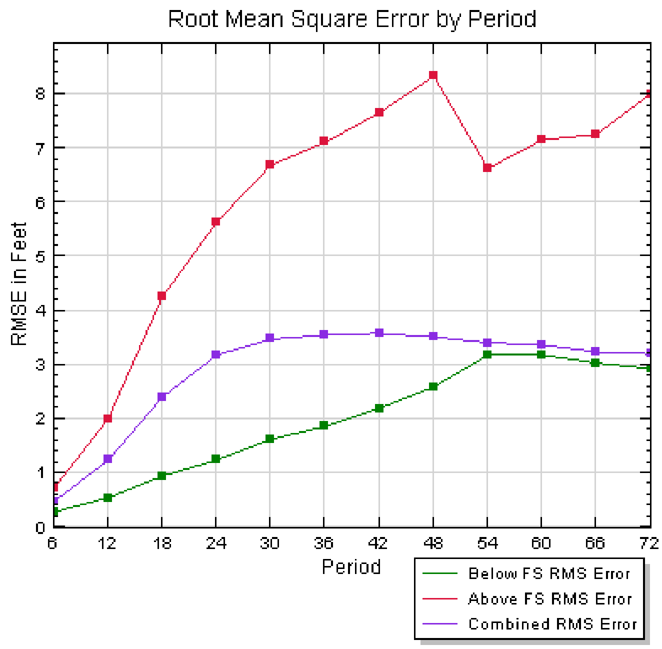
SERFC Verification Case Study
Presented on November 18, 2008 by Chris McGehee

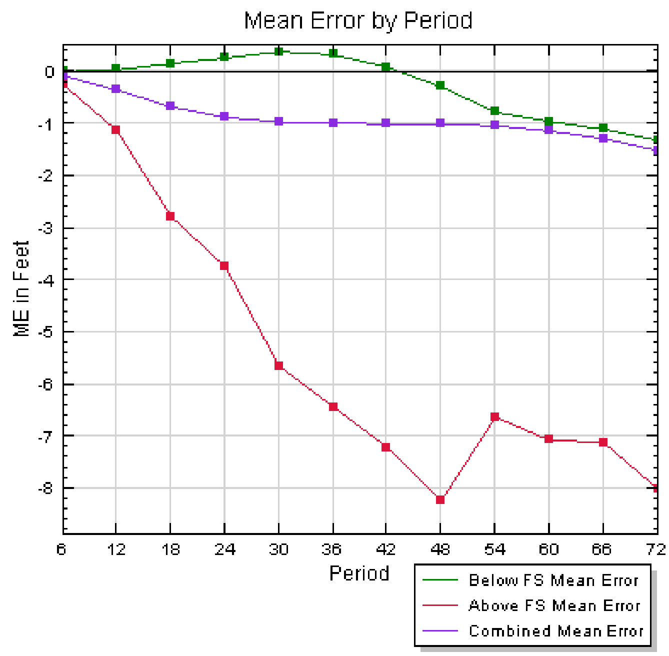
SERFC has been participating in the national statistical and Southern Region categorical river forecast verification programs since 2001. Here are some examples of the verification results for MKHF1 on the Manatee River near Myakka Head, FL, with some speculation as to their meaning. In the second part of this case study, we will present another possible verification method, with examples from this same site.

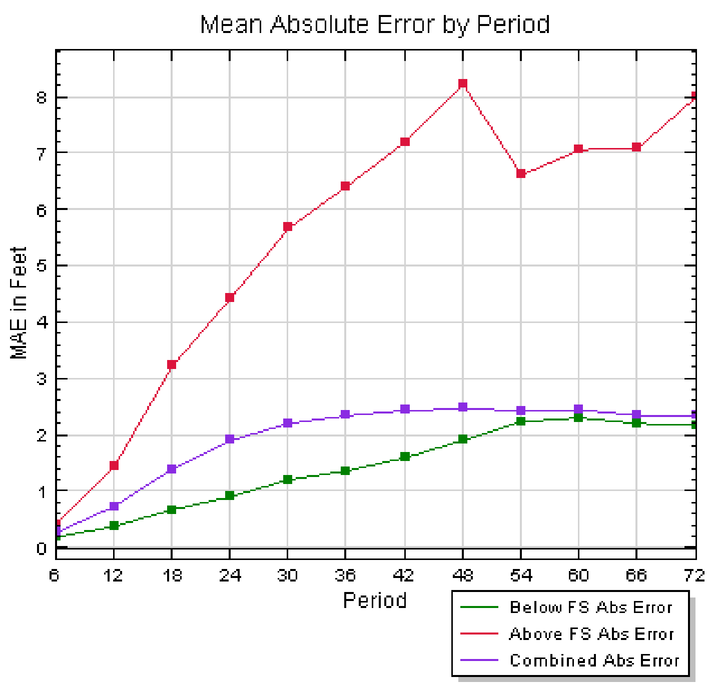
Verification analysis for MKHF1

MKHF1 is a flood-only forecast site, with a 65 square mile drainage area, and a time to crest that's typically from 18 to 30 hours. It is located in west central Florida on the Manatee River, a short distance above the pool of Lake Manatee. As with all SERFC forecast points, the MKHF1 forecast usually contains 24 hours of QPF. Per SERFC policy, if a tropical system is expected to impact the area, 60% of the Day 2 QPF is included as well. This is in an effort to account for uncertainty in storm track.

Using the river forecast verification stats-on-demand interface on the NWS Performance Management website, I generated statistics by forecast period since 2001. At MKHF1, root mean square error (RMSE), mean error (ME) and mean absolute error (MAE) show a tendency to start out with minimal error, worsen steadily through 48 hours, then improve markedly. Forecast error is substantially higher above flood stage than below.



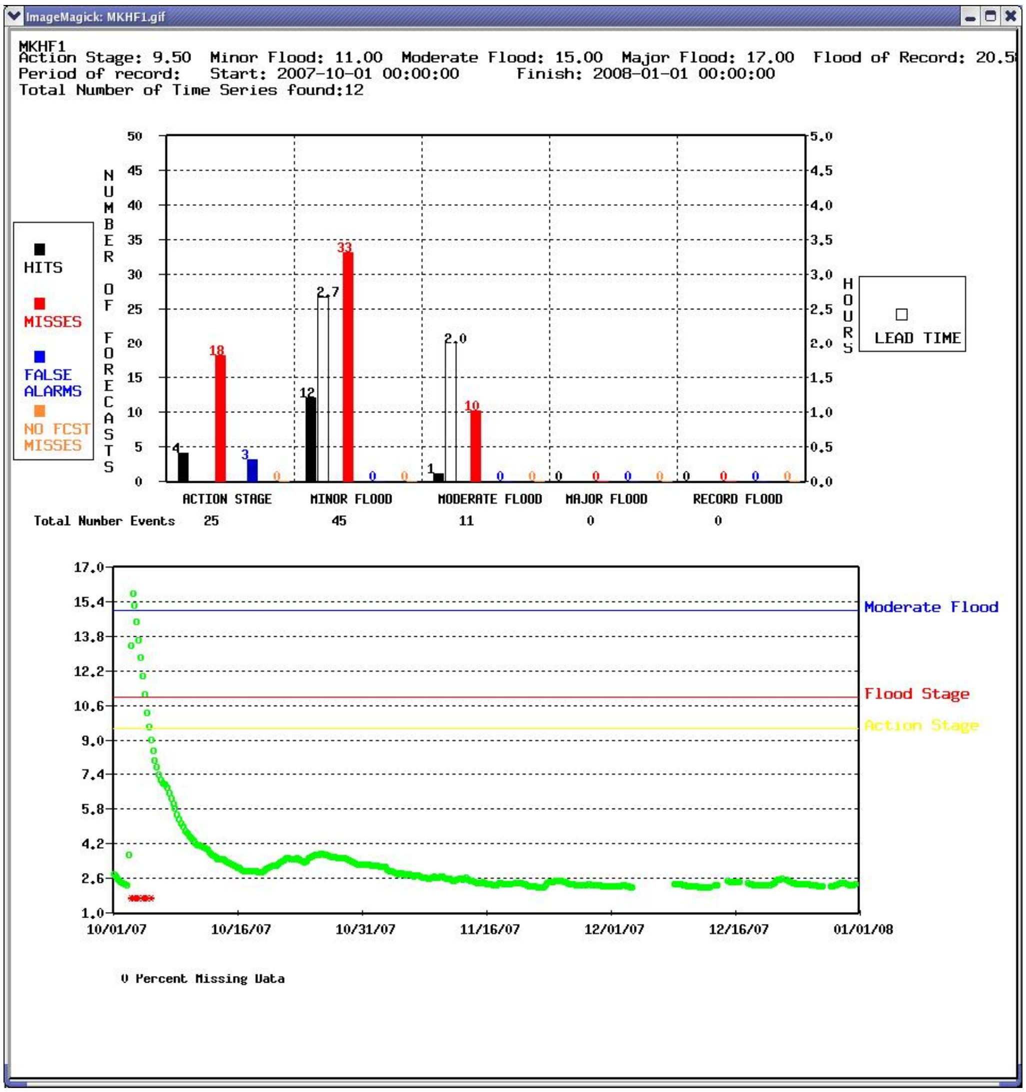




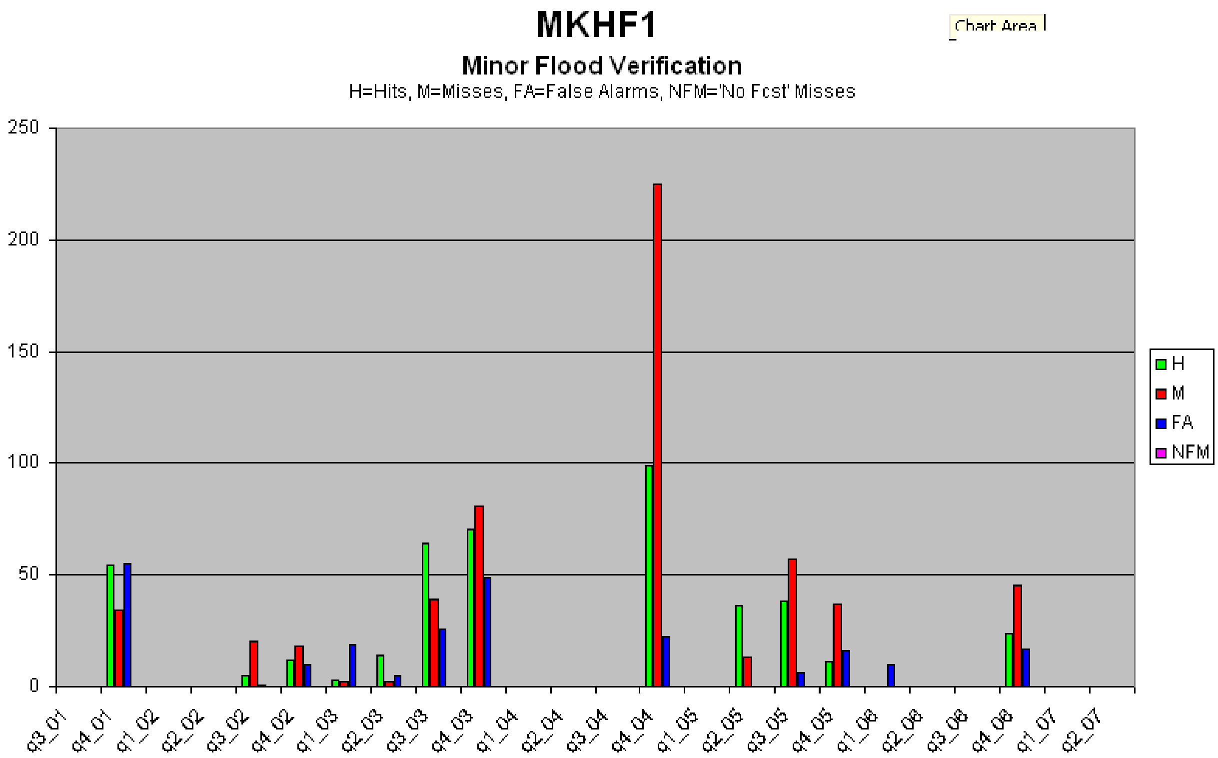
It should be noted that since this is a flood-only site, these statistics have been generated from events with significant precipitation; in addition, these events are mainly of the convective/tropical type, and are over a small basin. A daily forecast point would have lower-precipitation events and dry periods that would significantly improve the score. Therefore, it may not be appropriate to regard these results as an indictment of 24-hour QPF at this site, or of the 60% Day 2 policy.

At this time, SERFC only archives stage data and stage forecasts. It would be useful to run QPF-based river forecast verification for this and other sites; this would allow a more meaningful evaluation of the usefulness of 24-hour QPF as well as the Day 2 tropical policy. We hope to be able to do that in the future.

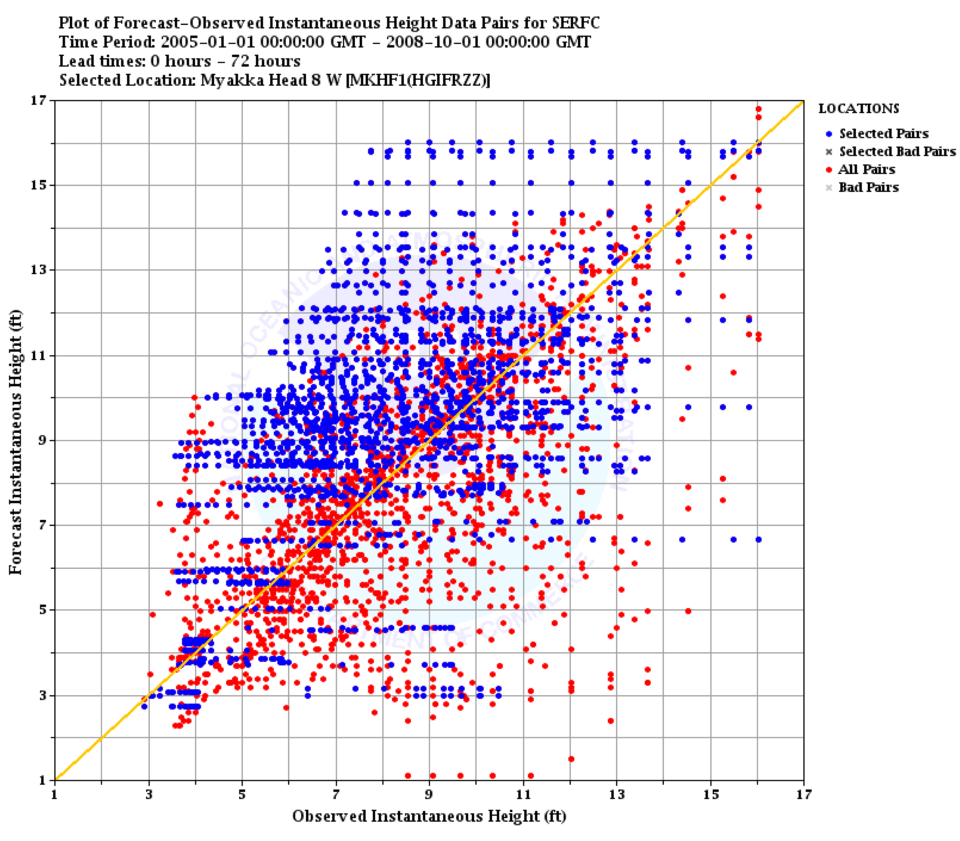
The Southern Region categorical verification program calculates statistics such as lead time, hits/misses and false alarms when sites are above action stage (or above minor flood stage if no action stage has been set for that site). Categories are action stage, minor/moderate/major and record flood. No verification statistics are generated when observations and forecasts are below action stage (or minor flood stage, as appropriate).



These statistics have been of little utility for providing insight on how to improve forecasts, as they only record whether a forecast was in the correct category, not if it was too high or too low. I aggregated the results over time, and divided them by forecast category, to see if we could derive more meaning from it that way.

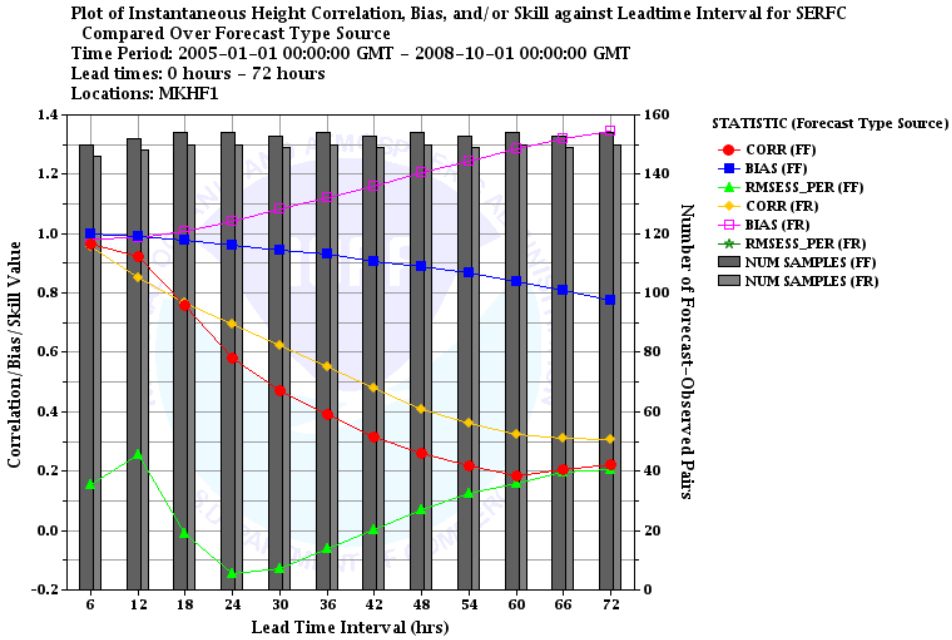


The new IVP software has a wealth of capabilities that will be useful for analyzing forecast quality. At present, SERFC only has forecast and persistence time series data to work with. In addition, due to several RAX failures, SERFC only has archived data from 2005 forward. The following is a plot of forecast and persistence pairs for MKHF1. The forecast pairs, in red, show a slight negative bias; the persistence pairs, in blue, show a slight positive bias.

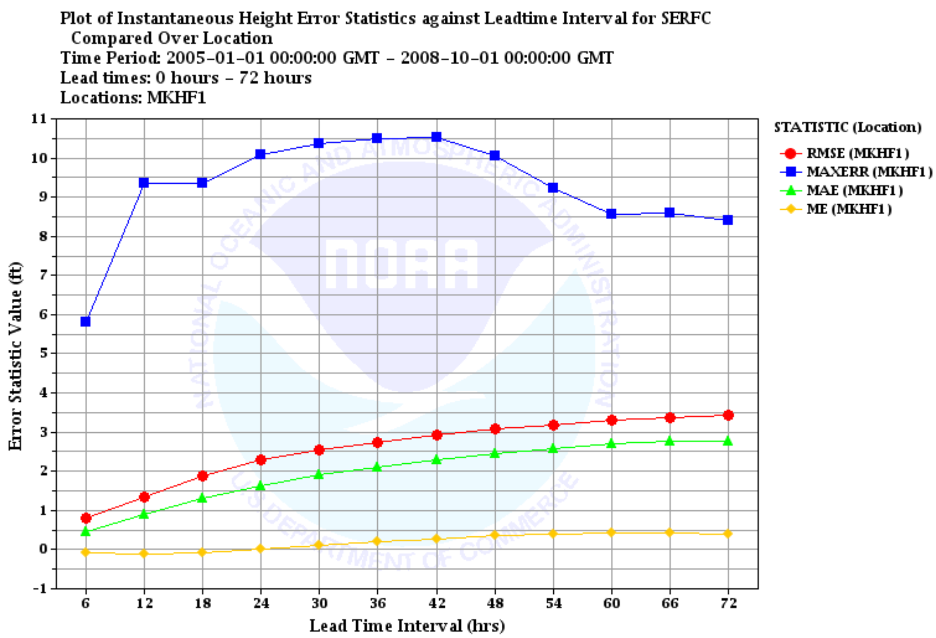


The following graphic shows an analysis of correlation, bias and skill for forecast and persistence. The forecasts show a small negative bias that increases slightly over time. Forecast correlation is good through 12 hours, then rapidly worsens through 48 hours, leveling off after that. RMSE skill score vs. persistence only shows skill in the first 12 hours.

.



The next graphic shows that, while MAE, ME and RMSE increase modestly over time, maximum error (MAXERR) jumps sharply between 6 and 12 hours lead time. MKHF1 can have a very steep rising limb when there is heavy rain, and this tendency is not well resolved by the model. The highest MAXERR is seen between 24 and 48 hours lead time. We would like to investigate this further when archived QPF data is available.



Forecast Hydrograph Feature Verification and Depiction: Proposal and Case Study (in collaboration with John Schmidt at SERFC)

Southern Region's categorical verification methods of deterministic hydrologic forecast time series have historically treated the members of the time series as discrete events with fixed lead times. One of the biggest influencing factors in that design was the difficulty in writing logic to analyze complex observed and forecast hydrographs to define an "event". One of the biggest liabilities of that design is that it does not calculate timing errors in forecast hydrographs in a meaningful way. That is, a sharp-crested forecast hydrograph that accurately predicts the crest height but is too early or late by six hours can be graded very poorly while a forecast hydrograph that is too high or low by a couple of feet but accurately predicts the timing of the crest may grade well in comparison. This scenario is depicted in Figures 1 and 2.

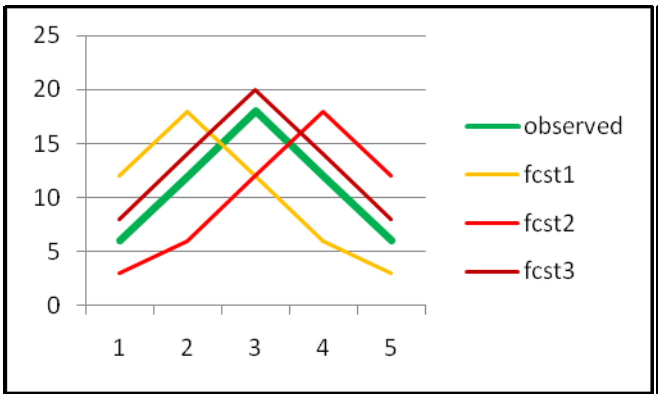


Figure 1. 3 "dummy" forecasts plotted against "observed".

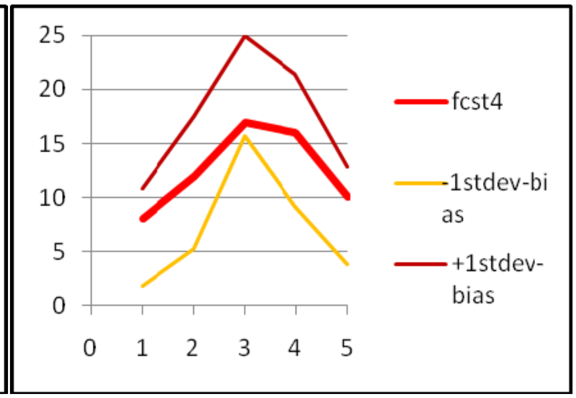


Figure 2. Depicting legacy lead time-based error on a forecast hydrograph.

If we use the alternative verification logic on the three forecast hydrographs in Figure 1 and calculate Average Error for each lead time what we see is a high bias at Time 1, a high bias at Time 2, a low bias at Time 3, a high bias at Time 4 and a high bias at Time5. If we further calculate a Standard Deviation for each lead time and then use the "bias" calculated previously and that Standard Deviation as a reasonable maximum and minimum condition and apply it to a subsequent forecast, fcst4, we can depict these statistics as in Figure 2. Is this accurate? Is this meaningful? Is it describing our skill at forecasting at this location?

Our proposal and case study analyze each hydrograph as a single forecast, verifying the error in forecast crest timing and magnitude independently. Using that type of logic on the forecasts in Figure 1 we see that fcst1 had no error in magnitude, but was 1 time step early. Fcst2 was 2 feet high, but had no error in timing and fcst3 had no error in magnitude, but was 1 time step late. If we calculate the Average Error and then Standard Deviation on these 3 crests we see no "bias" in timing and a slight "high bias" in magnitude. We can depict this verification analysis on the same fcst4 hydrograph by centering an ellipse on the calculated timing and magnitude bias and extending it to the calculated Standard Deviations. This is depicted in Figure 3, along with the "error" depicted in Figure 2. Is this accurate? Is this meaningful? Is it describing our skill at forecasting at this location?

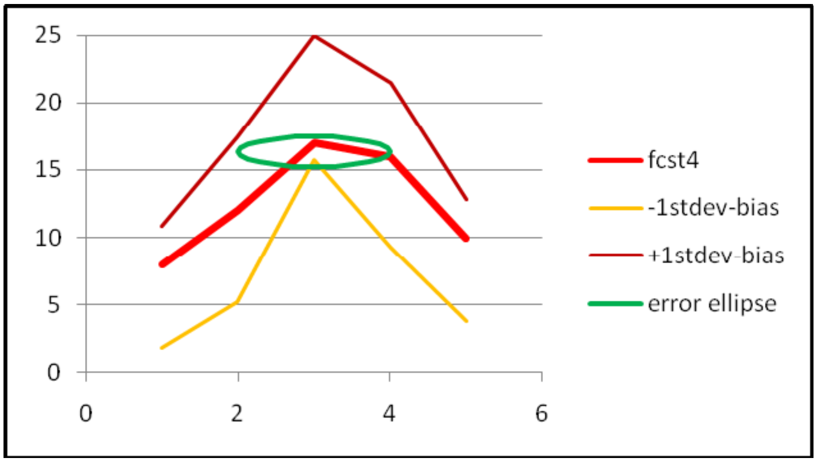


Figure 3. Lead-time based verification vs. crest-based verification.

How would this work with real data? SERFC has a local archive of RVFs at Myakka Head, FL from Feb. 2005 to current. Although much of SERFC has been under a historic drought during most of that time, the Myakka Head forecast point has had 17 observed crests above flood stage. The SERFC issued 47 forecasts before crest during those events. We selected only the 43 forecasts of Minor Flooding for this case study. They are listed along with the simple verification statistics in Table 1.

MINOR	fcst lead time	max height	max time	ob max	ob time	ob error	time error
5/4/2005	12.27	11	16	10.59	5	0.41	11
6/30/2005	16.52	11	6	13.19	14	-2.19	-8
7/17/2005	10.75	11	0	11.41	16	-0.41	-16
6/6/2005	15.58	11.2	30	10.37	15	0.83	15
6/2/2005	22.07	11.3	12	8.38	2	2.92	10
6/30/2005	5.88	11.3	6	13.19	14	-1.89	-8
7/18/2005	4.33	11.4	18	11.41	16	-0.01	2
2/28/2005	10.48	11.5	6	16.17	8	-4.67	-2
6/29/2005	3.37	11.5	18	11.39	16	0.11	2
7/10/2005	27.88	11.5	18	9.5	7	2	11
10/03/2007	9.5	11.5	12	15.82	12	-4.32	0
6/29/2005	9.23	11.6	12	11.39	16	0.21	-4
6/10/2005	45.07	11.7	36	12.62	23	-0.92	13
7/8/2005	56.5	11.7	0	9.5	7	2.2	-7
6/30/2005	3.42	11.9	6	13.19	14	-1.29	-8
8/9/2005	5.53	12	18	12	19	0	-1
8/9/2005	4.37	12	18	12	19	0	-1
10/25/2005	22.35	12	36	10.48	19	1.52	17
3/18/2005	14.13	12.1	6	13.71	9	-1.61	-3
6/10/2005	39.05	12.2	36	12.62	23	-0.42	13
7/9/2005	32.5	12.2	0	9.5	7	2.7	-7
9/20/2006	10.62	12.2	0	13.44	4	-1.24	-4
6/22/2008	7.5	12.3	18	12.4	19	-0.1	-1
3/18/2005	8.88	12.4	6	13.71	9	-1.31	-3
7/1/2005	26.93	12.4	42	11.7	16	0.7	26
6/22/2008	6.9	12.4	21	12.4	19	0	2

6/11/2005	8.95	13	24	12.62	23	0.38	1
10/25/2005	16.17	13	18	10.48	19	2.52	-1
10/3/2007	2.37	13.1	6	15.82	12	-2.72	-6
10/23/2005	33.92	13.4	0	10.48	19	2.92	-19
10/24/2005	21.97	13.4	0	10.48	19	2.92	-19
10/24/2005	9.95	13.4	0	10.48	19	2.92	-19
6/30/2005	11.2	13.5	18	13.19	14	0.31	4
7/1/2005	4.03	13.5	18	13.19	14	0.31	4
9/17/2006	12.93	13.5	24	14.36	17	-0.86	7
9/21/2006	11.02	13.6	12	13.44	4	0.16	8
3/18/2005	21.07	13.7	12	13.71	9	-0.01	3
10/3/2007	0.17	13.8	6	15.82	12	-2.02	-6
10/3/2007	5.48	13.8	12	15.82	12	-2.02	0
6/11/2005	52.4	14.3	54	12.62	23	1.68	31
9/17/2006	16.08	14.4	30	14.36	17	0.04	13
9/17/2006	7.68	14.5	24	14.36	17	0.14	7
2/28/2005	14.37	14.9	12	16.17	8	-1.27	4
					AVG		
					ERR	-0.03209302	1.418604651
					ST DEV	1.818707754	10.80615468
					MAX	2.92	31
					MIN	-4.67	-19
					2 st dev	3.637415509	21.61230936

Table 1. 43 Myakka Head, FL minor flood forecasts verified.

When these forecasts were analyzed en masse, no significant bias in forecast crest magnitude is noted and a slight +1.4 hour bias in timing was calculated indicating the forecast crests were a little later than observed. The observed crest height and time values were pulled from hourly data and the forecast values are based on a six hour time step.

Figure 4 shows the application of the Average Error and Standard Deviation to a subsequent “dummy” Myakka Head forecast in the form of an ellipse. The yellow-shaded area is Action Stage. The orange-shaded area is Minor Flood Stage and the red-shaded area is Moderate Flood. The green circles represent six-hourly “observed” data. The pink “X”s depict a minor flood forecast hydrograph with six-hourly ordinates. The two plotted ellipses are calculated using the real data in Table 1. The smaller ellipse depicts the area that, based on our past 3-year history, 67% (1 standard deviation) of observed crests should fall in when we forecast a minor flood crest. The larger ellipse depicts 2 standard deviations or about 95% of events. Similar verification calculations could also be done on other forecast hydrograph features such as “Rise Above XXXX Flood Category” and “Fall Below XXXXX Flood Category”.

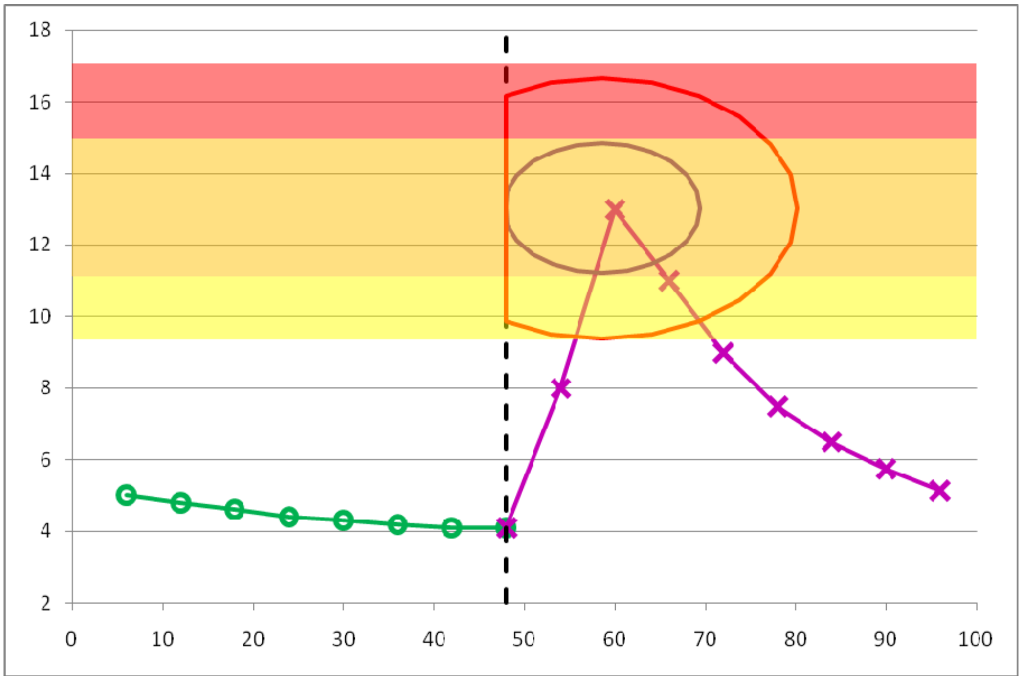


Figure 4. Depiction of 1 and 2 standard deviation(s) of error in timing and magnitude for Myakka Head, FL minor flood forecasts.

Verification of hydrologic forecasts serves many purposes. The previous example describes a method of depicting overall forecast skill to users. Another purpose is identifying systematic biases to allow for modifications in the forecast process. With that in mind, we subdivided these forecasts into ranges of lead time and reran the same statistical analysis on these subsets. Figure 5 shows the same 1 standard deviation ellipse, centered on the bias, of minor flood forecasts with a lead time of 12 hours or less, a lead time between 12 and 24 hours and finally a lead time in excess of 24 hours.

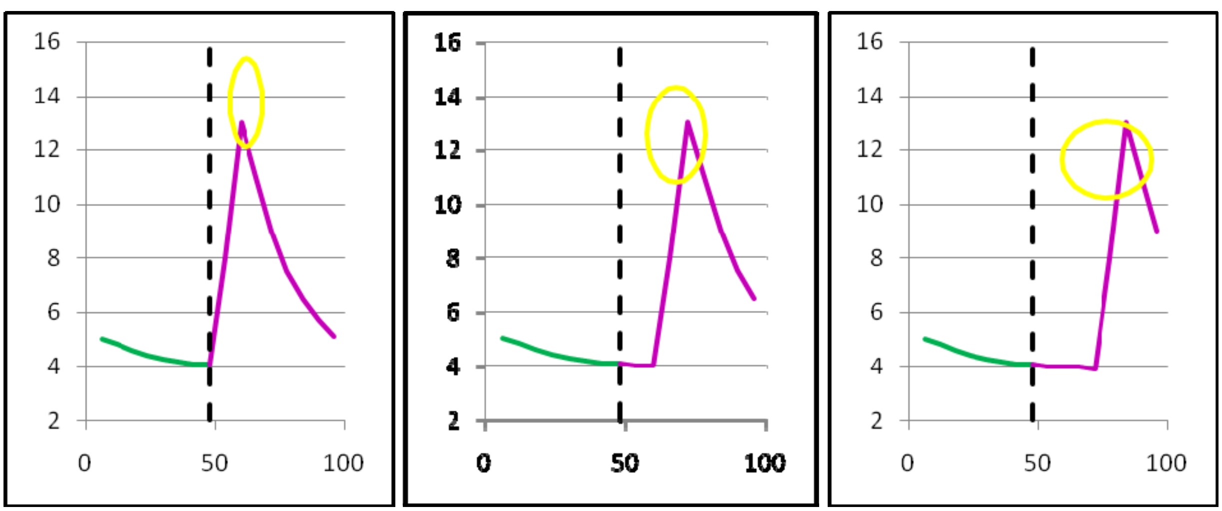


Figure 5. Error ellipse for <12 hour lead time, 12-24 lead time and >24 hour lead time minor flood forecasts.

When these events are subdivided we do start to see some biases. In the short-term window we are early and low in forecasting minor flood events. In the intermediate-term window we are a little high and late. In the long-term window we are high and late. So, one question is how useful is this as a forecasting tool? Should forecasters modify their forecasts based on their historical biases? Does that make the depiction of the historical invalid then?

There are other issues including the following ones:

- Is there some logic to automate the event definitions in the observed data? In most cases forecast hydrographs only have one crest, but not always.
- Is this depiction of the error accurate?
- Should it be a rectangle instead of an ellipse?
- Is there a dependence/independence issue we're ignoring?

NERFC Verification Case Study
Presented on September 03, 2008 by Rob Shedd

Background

For the past several years, Eastern Region has instituted what they refer to as GPRA+ measures for the offices in the region. One of the RFC GPRA+ goals is that the RFC QPF on a monthly basis should improve upon HPC QPF for the 0.5 to 1.0 inch Day 1 QPF. Over the past couple of years, NERFC performance on this standard has been slipping. In response, the NERFC has instituted several projects to gain a better understanding of QPF performance and to attempt to assess whether the errors are more significant in certain forecast basins, certain storm types, and whether there is a significant difference between individual forecasters. The desire was to perform this assessment over QPF basins and comparing to operational QPF, rather than the NPVU approach of verifying QPF grids against MAPX grids which are not always used operationally.

Requirements

Several tasks were initially required in order to prepare data for this analysis. First, the HPC QPF grids had to be converted to basin QPF. A procedure was put in place to daily download the HPC grid formatted QPF (in addition to the vgf files which we routinely receive). These grids are then converted to a QPS formatted SHEF product using GRASS GIS. A text product is made available to the forecasters, and is also posted to the RFC archive database. HPC data is posted to the database with a type source code of FP. Similarly, NDFD grids, although not specifically used in this report yet, are also converted in the same way and posted to the database with a type source code of FN.

Finally, the RFC QPS products are also posted to the archive database with a type source coce of FM. In addition, when NMAP QPF is run, a log is made of the individual forecaster who prepared the QPF. This data is accessible only by the DOH (under an agreement with the local NWSEO steward). Duplicates of the RFC QPF records are then made in the archive database with a type source code that is unique for each individual forecaster.

The existing verification system pairing program would not allow the type of pairing that was required to copy the necessary data into the vsys_pairs tables for use by the IVP. As a result, a separate program was required to perform this pairing operation.

MAP and MAPX data from NWSRFS OFS files was also posted to the archive database using the ofsshef routines available on the archive system. For this project, the MAP data was the comparison data set.

Available Data

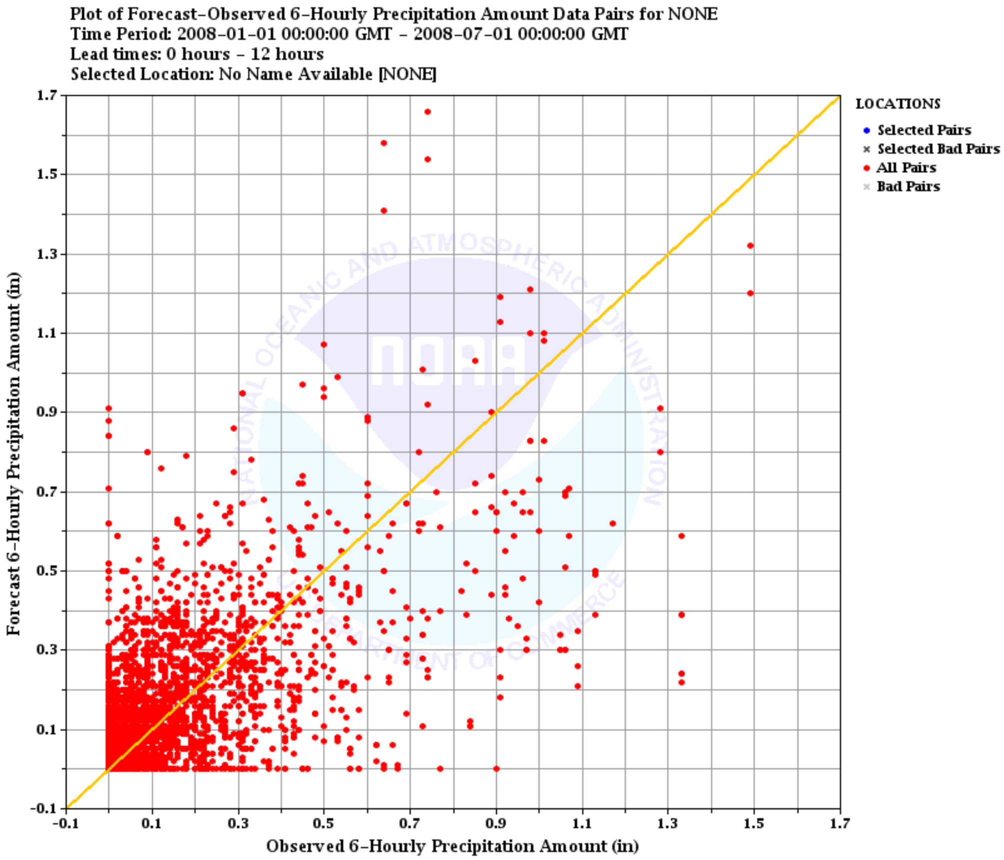
A complete set of data for the six month period, January – June 2008, was available that included RFC, HPC, and individual forecaster pairs. For a number of reasons, this resulted in a very limited demonstration of the capabilities of the system.

- We were not initially aware of the capabilities for IVP to increase the available memory for its runs. As a result, the amount of data that could be accessed at a single time was very limited. However, this has since been resolved, although time constraints have not yet allowed us to perform updated analysis on a more comprehensive set of data. As a result, most of the analysis in this report is focused on limited set of 10 forecast basins of various size and spread throughout our forecast region.
- Having only six months of data available, resulted in a fairly small number of samples of more significant rainfall amounts of greater than half an inch which was the primary focus of the study.
- Similarly with each forecaster working only a couple of sets of operational HAS forecasters during the time, the sample size for individual forecaster data was even smaller

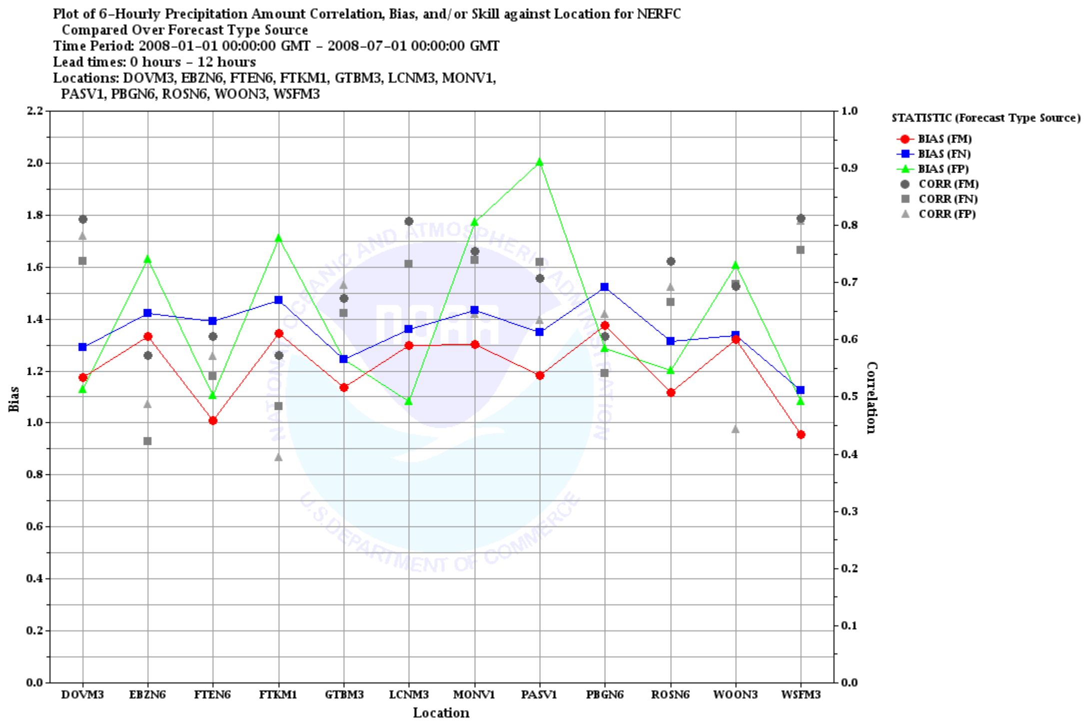
As a result of these limitations, this study was more of a proof of concept than a complete verification study. A more comprehensive study will be performed in the near future.

IVP Analysis

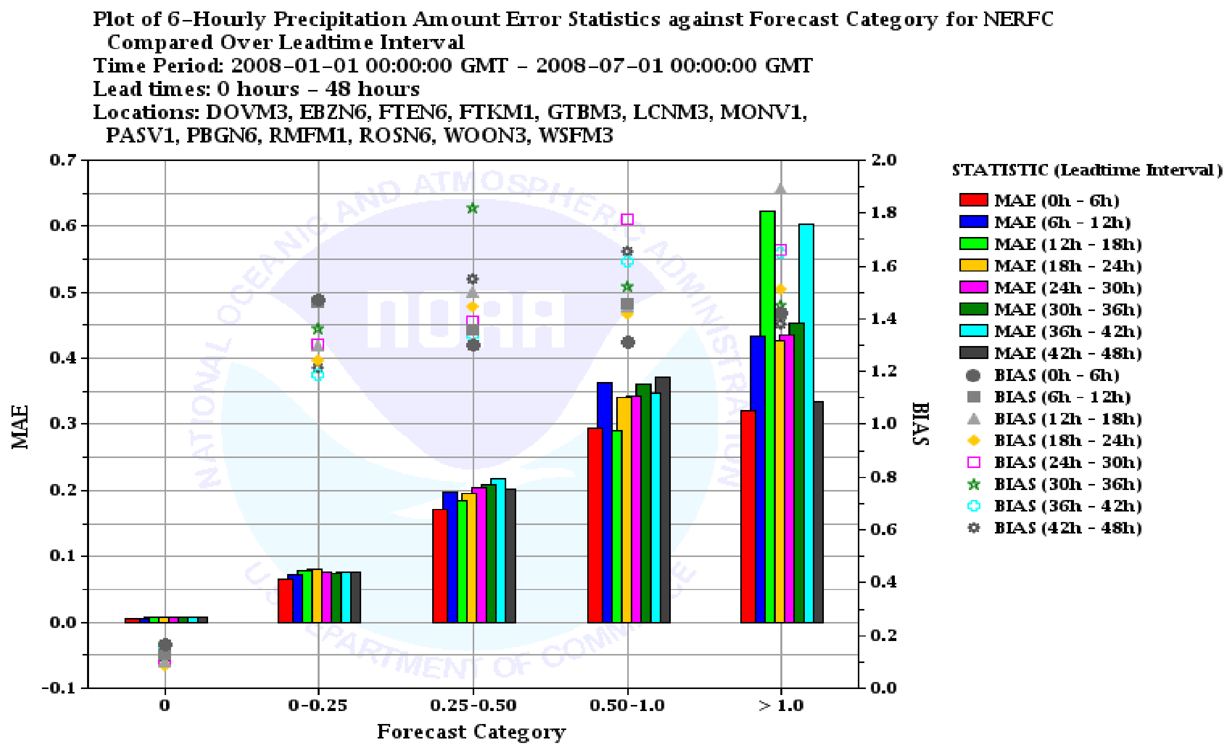
The initial comparison plot for this project was the basic scatter plot comparing all observed vs. forecast data. While data is generally centered on the x=y line, there is a great deal of scatter. This also highlights the small sample size at the higher amounts.



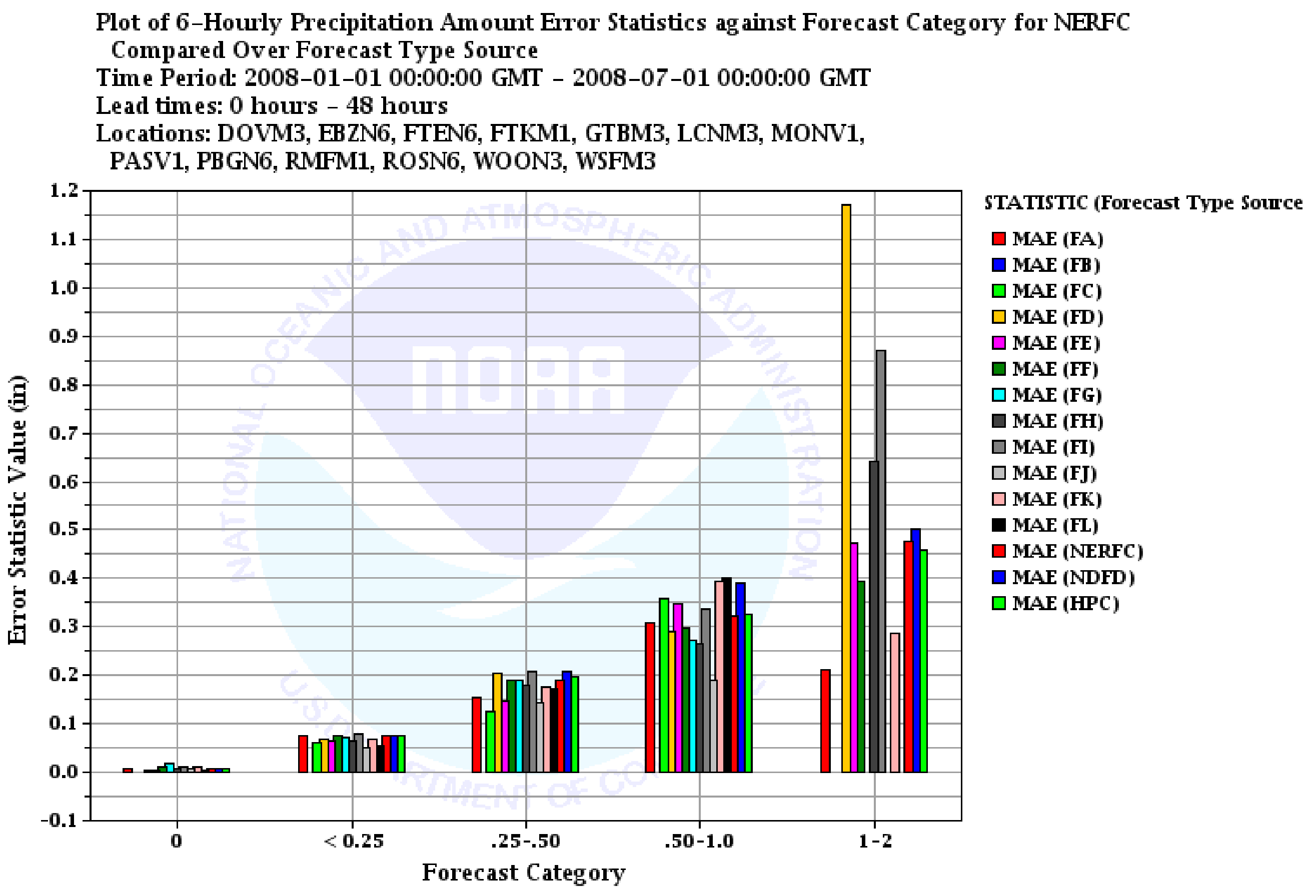
Another plot that was looked at was a comparison of the bias and correlation at several forecast points and separated by forecast source. The following plot makes clear that in the 0-12 hour period which NERFC focuses on, NERFC forecasts generally have a much better bias and correlation than either the HPC or NDFD forecasts. However, looking at a similar plot for the 0-24 hour period (not shown) shows much less improvement by NERFC.



Another plot performed a comparison of Mean Absolute Error by Lead time and forecast category. This clearly shows that MAE increases both with lead time and forecast category as one would normally expect.



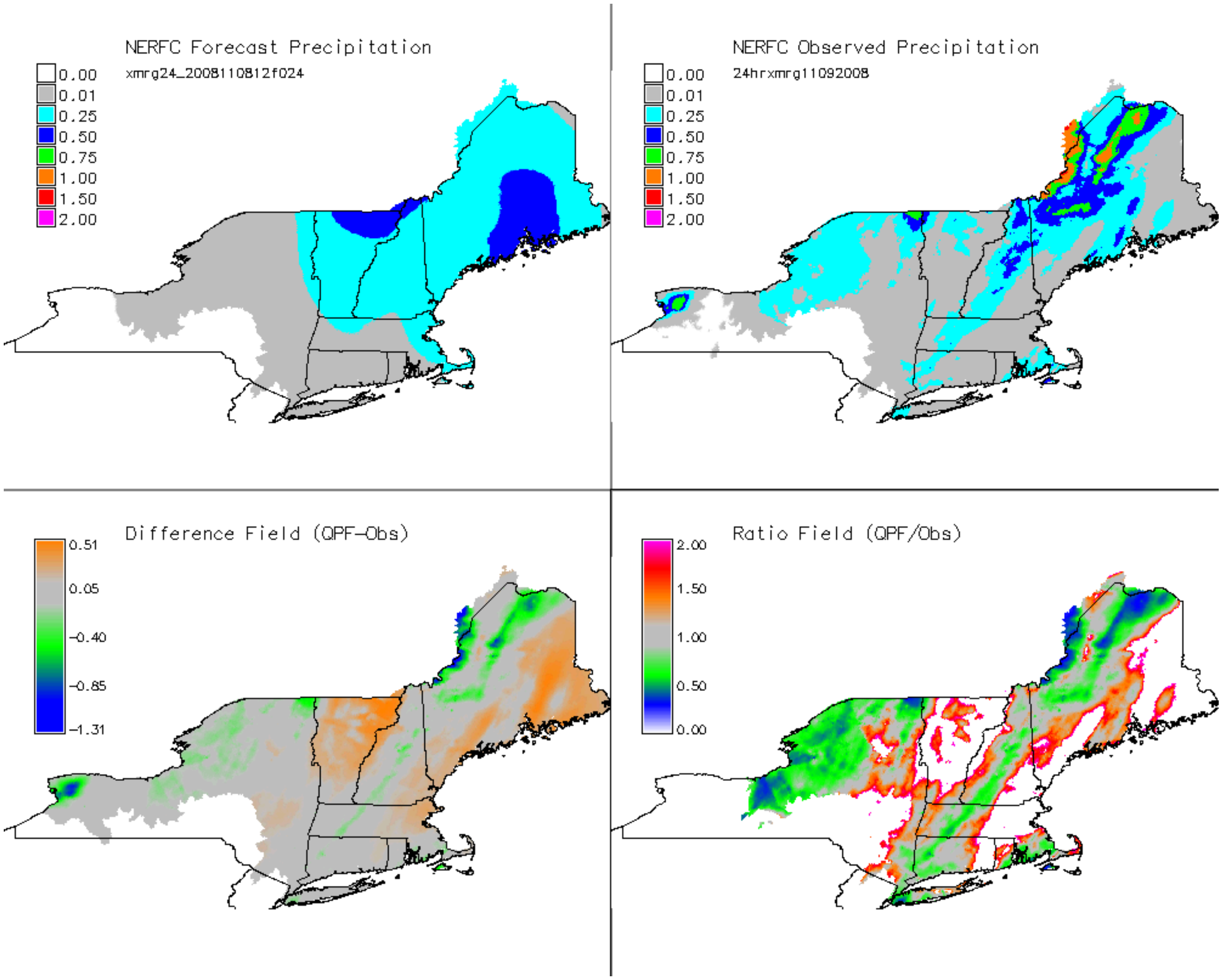
The final plot below demonstrates a display of the individual forecaster performance. This demonstrates some difference between forecasters; however, given that the sample size in some of the higher category cases was very limited, no conclusions should be drawn from this plot at this time.



Summary

NERFC plans to continue our work investigating QPF performance during the current year. We would like to develop a mechanism that allows an individual forecaster to be able to go in at any time and see his/her performance relative to the staff at large and HPC. Immediate feedback is also desirable to support this activity to assess their performance yesterday. As such, we have developed a graphic that provides a comparison of QPF grids on a 6-hr and 24-hr basis relative to MAPX grids. These are being computed continually to provide more immediate use to the forecaster.

In addition, we hope to expand these comparisons to handle a longer time period, more forecast locations as well as provide comparisons of MAP vs. MAPX. As we move into a more gridded world, we need to be able to understand the performance of the different model forcings that are being used.


NERFC Forecast Precipitation
xmrg24_2008110812f024
0.00
0.01
0.25
0.50
0.75
1.00
1.50
2.00
NERFC Observed Precipitation
24hrxmrg11092008
0.00
0.01
0.25
0.50
0.75
1.00
1.50
2.00
Difference Field (QPF-Obs)
0.51
0.05
-0.40
-0.85
-1.31
Ratio Field (QPF/Obs)
2.00
1.50
1.00
0.50
0.00

OHRFC Verification Case Study
Presented on September 25, 2008 by Tom Adams

Introduction

The OHRFC Verification Case Study is aimed at identifying:

(1) the value of Quantitative Precipitation Forecasts (QPF) compared to ZERO QPF;
(2) the value of runtime modifications (MODs) versus no use of MODs;
(3) the superiority of OHRFC Hydrometeorological Analysis & Support (HAS) QPF compared to Hydrometeorological Prediction Center (HPC) QPF, if any exists;
(4) verification differences across forecast points.

Item (3) is particularly important because there is skepticism that RFC HAS units can outperform HPC QPF and because the OHRFC feels that HPC QPF significantly over estimates heavy precipitation for hydrologic forecasting. The study was made for only a small number of basins that differed both geographically and in basin area.

Methodology

Six parallel, non-operational, batch Operational Forecast System (OFS) runs were made *once daily*, following saving carryover after the morning operational forecast run. Consequently, all model runs began with the same initial Sacramento Soil Moisture Accounting (SAC-SMA), SNOW-17, and routing model conditions at the beginning of the model run period (the OFS STARTDATE). Seven basins were selected to cover a wide range geographically and by basin area. Hence, a widely varying range of hydrologic response was evaluated. The following batch OFS runs were made:

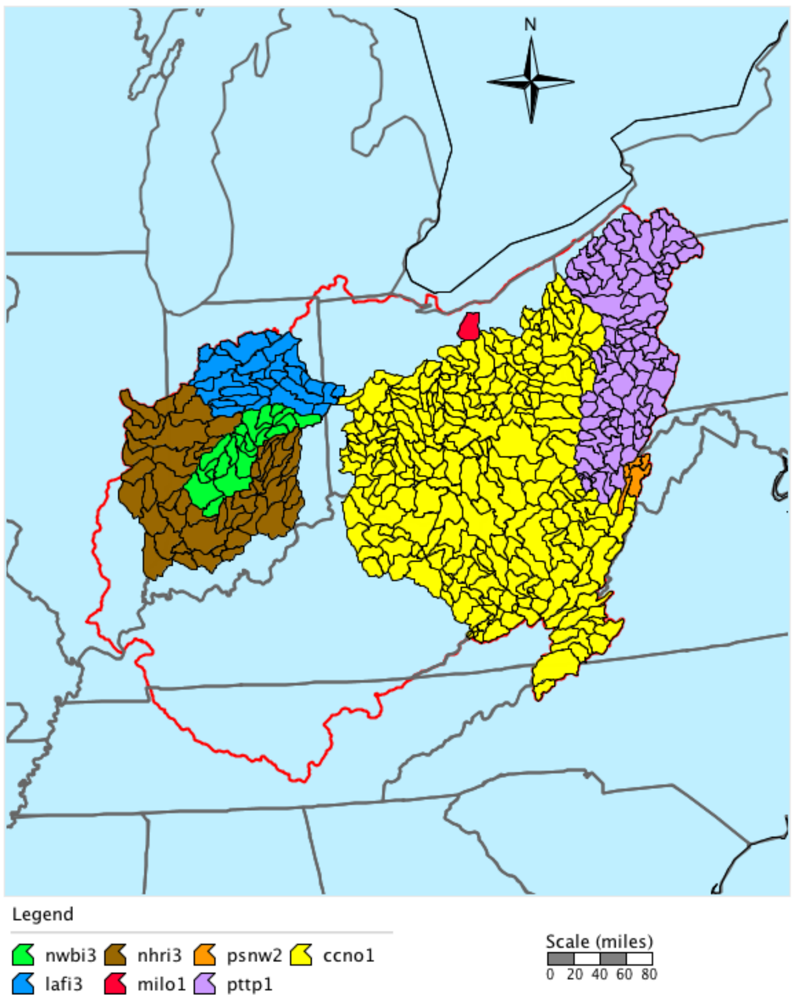
- no MODs, no QPF
- no MODS, with HPC QPF
- no MODs, with HAS QPF
- with MODs, no QPF
- with MODs, with HPC QPF
- with MODs, with HAS QPF

Additionally, OHRFC operational forecast runs were used in the study[1] and served as a basis to investigate the influence of differing sample size, since some of the basins studied have at least two daily forecast runs due to a routine evening update. This doubles the sample size for some basins. However, two basins are flood only[2] (non-daily) forecast points, so the number of operational forecasts is much smaller than the once-daily experimental runs. Following the batch OFS model runs the PRDUTIL TSDATA command was run to dump the time series data into ascii files. Subsequently, a custom Perl script was used to reformat the files into a format suitable

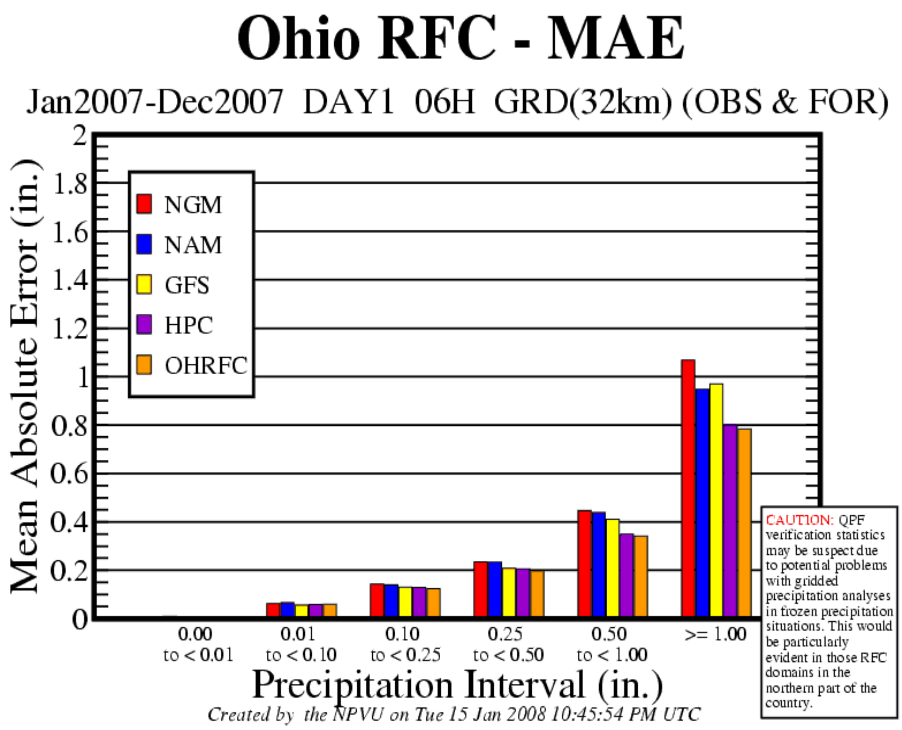
[1] The operational runs were archived directly, rather than using the PRDUTIL TSDATA command.
[2] Forecasts are issued if forecast stages exceed *Action Stage* for these locations.

for importing the data into the PostgreSQL Archive Database. Once the data was successfully imported into the Archive database, the *Interactive Verification Program* (IVP) was used to analyze the data and to generate graphical plots.

The map shows the location of the basins used in the study including the individual Mean Areal Precipitation (MAP) areas. Note that all upstream basins contribute flow to the larger, color coded) downstream regions, so, strictly speaking, the verification results are not completely independent due to the upstream contributions.



The graphic below (Ohio RFC - MAE) is representative of the NPVU QPF verification statistics showing that OHRFC HAS results are equal to or better than HPC results. While the results show QPF improvement over the HPC QPF, the difference may not be statistically significant.

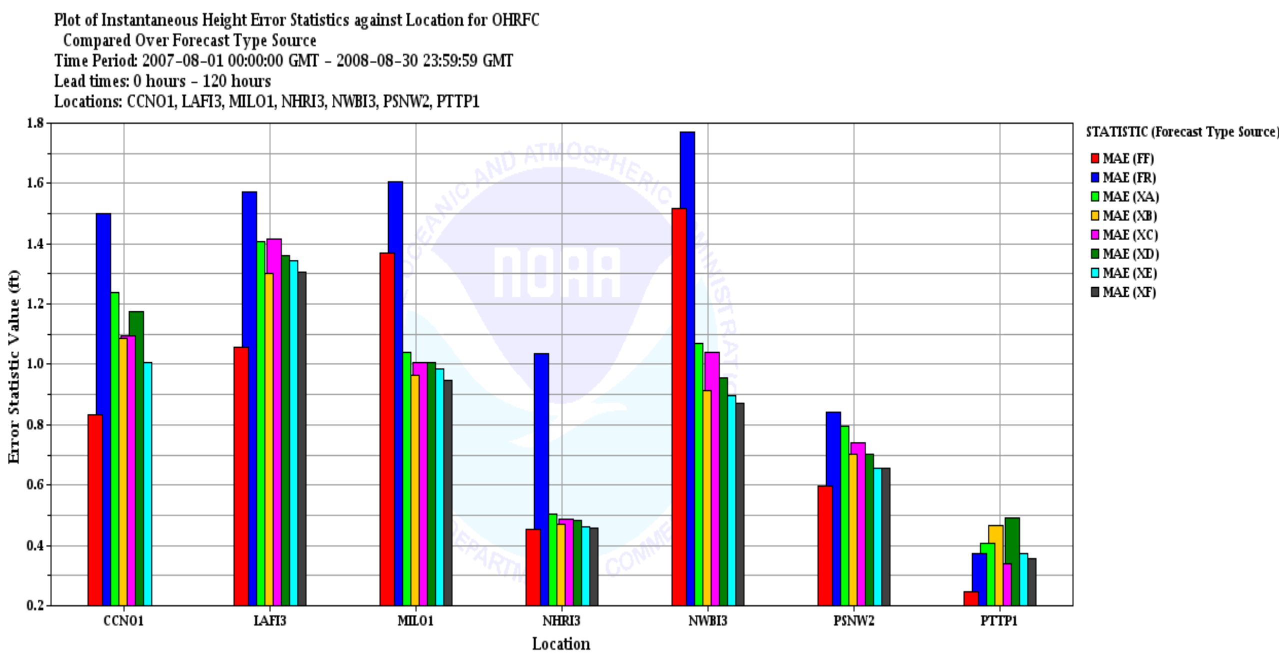


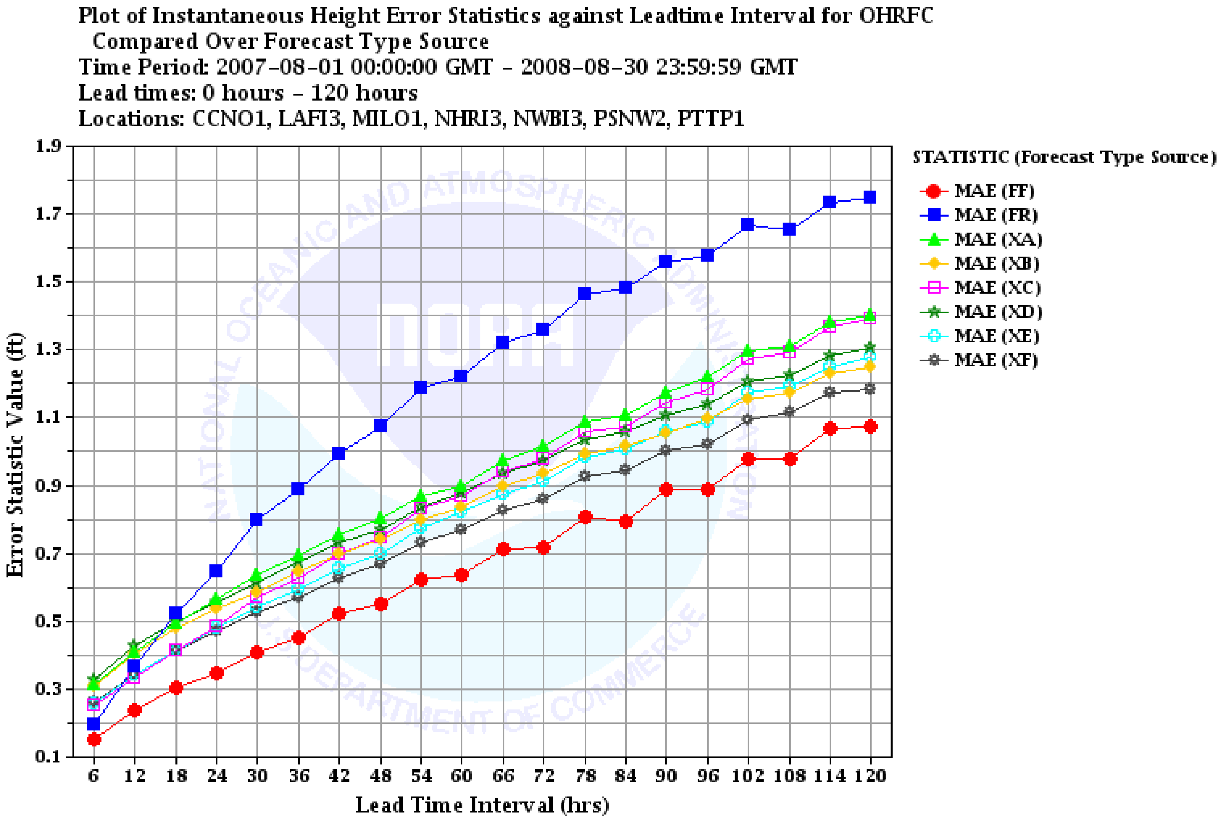
Below are two tables summarizing the different forecast points, along with their response time category and basin are, and the 8 sets of forecasts inter-compared in this case study.

ID	Forecast Group	Response	Basin Area (mi^2)
MILO1	GTL	fast	371
PSNW2	MNU	fast	722
NWBI3	WHT	medium	4688
LAFI3	WBU	medium	7267
PTTP1	OHW	medium	19101
NHRI3	WBL	slow	29234
CCNO1	OHC	slow	76580

Code	Meaning
XA	No MODs, No QPF
XB	No MODs, with QPF
XC	with MODs, No QPF
XD	No MODs, with HPC
XE	with MODs, with HPC
XF	with MODs, with QPF
FF	Operational Forecast
FR	Persistence

Some example verification graphics are shown below for Mean Absolute Error (MAE) grouped by forecast point, across all lead-times and by lead-time, grouping all forecast points. Many other analyses and graphics were produced, but are not included due to space limitations.





Conclusions

(1) The use of QPF improves forecast over ZERO QPF; this is more apparent for larger basins.
(2) The use of OFS run-time Modifications (MODs) generally improves forecasts; for large, slow-response and very fast responding basins, the use of MODs provides little or no improvements over no MODs.
(3) OHRFC hydrologic forecasts are better with the use of OHRFC HAS produced QPF over using HPC QPF. This improvement does not appear to be large.
(4) Careful scrutiny must be applied before drawing conclusions from verification results due to the influence of factors such as the influence of lock & dam operations, comparisons between basins of greatly differing sizes, and poor model calibrations (which should be evident).
(5) Statistics worse for flood-only points due to small sample sizes when compared to once daily experimental forecast runs.
(6) Sample size can greatly influence the verification statistics; this fact should not be under-estimated.

The OHRFC has begun the experimental process to evaluate the influence of QPFs of varying durations, namely, using HPC QPF for 6-, 12-, 24, 36-, 48-, and 72-hours. The OHRFC will continue with analyzing the verification methods described in this study for all remaining forecast points.

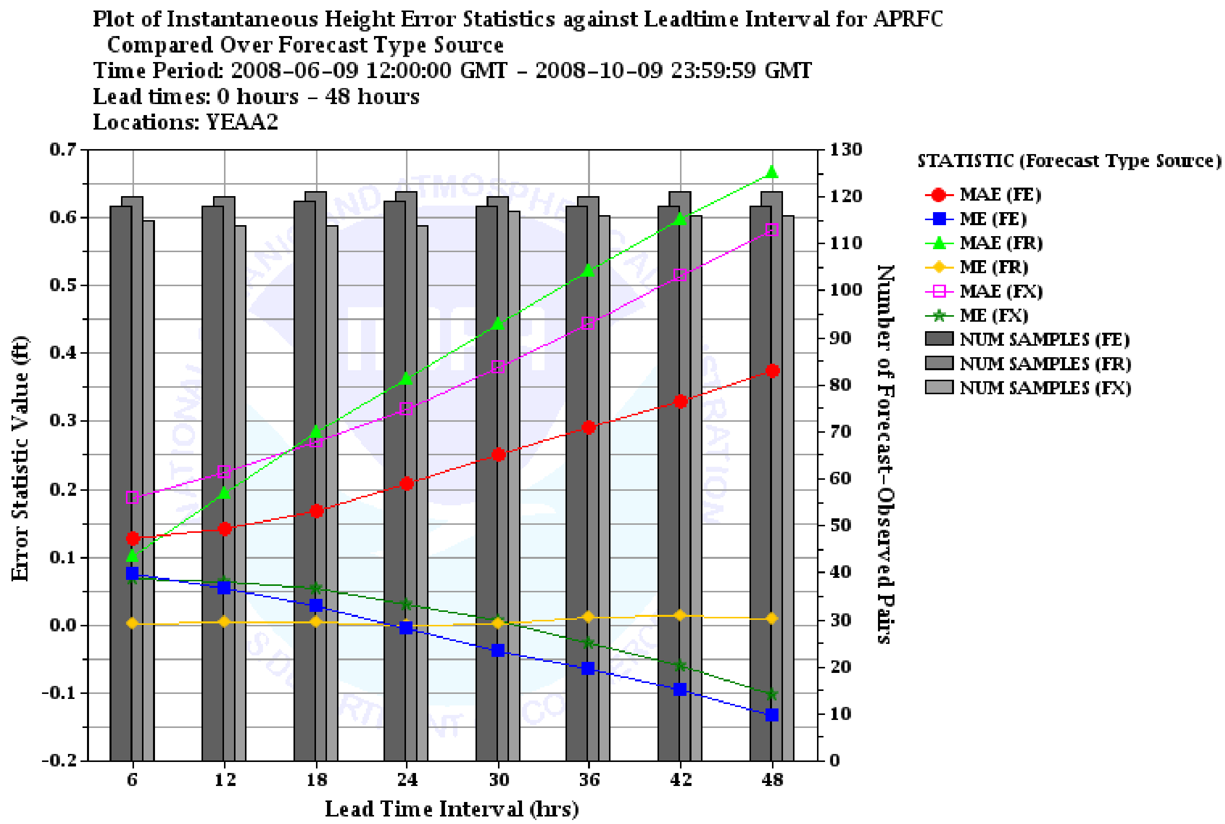
APRFC Verification Case Study
Presented on November 18, 2008 by Jim Coe

During the development season of 11/2007 to 4/2008, one of the major projects undertaken was to calibrate the upper Yukon River basins in Canada using model reanalysis data. The drainage area above the first forecast point in Alaska (Eagle – YEAA2) constitutes approximately 102,000 square miles or about 30% of the Yukon's total drainage area. Prior to this calibration effort the only calibrated basins being used to forecast at Eagle were Mayo (SRFQ9) and Stewart River at Mouth (SRMQ9). The remaining upstream points in Canada were used only as flow routing points and future flows for them were produced by a simple recession. No soil moisture or snow modeling was done and local inflows to Dawson (YDAQ9) and Eagle were computed as a percentage of flow at 'representative' locations. Since no significant modeling was taking place, the only mods used were generally flow time series mods.

For the new calibrations, model reanalysis data was used to create historical precipitation, while data from Environment Canada was used for historical temperatures. Reanalysis points were chosen that corresponded closely to current GFS model points in location and elevation. The new segments were implemented into NWSRFS in April 2008 and, at the same time, the old method of modeling at Eagle was made into a separate forecast group that would run on a cron with only data QC affecting the forecast from the old model. QPF is not currently used in either model.

APRFC decided that a verification case study to compare the newly calibrated basin to the old forecast method and to persistence would be a valuable exercise. We hoped to use the IVP program to show that the calibration effort would provide better forecasts at Eagle – particularly into the second day. Fortunately, the results indicate that this is so and have encouraged us to continue using reanalysis data for calibration in data sparse areas.

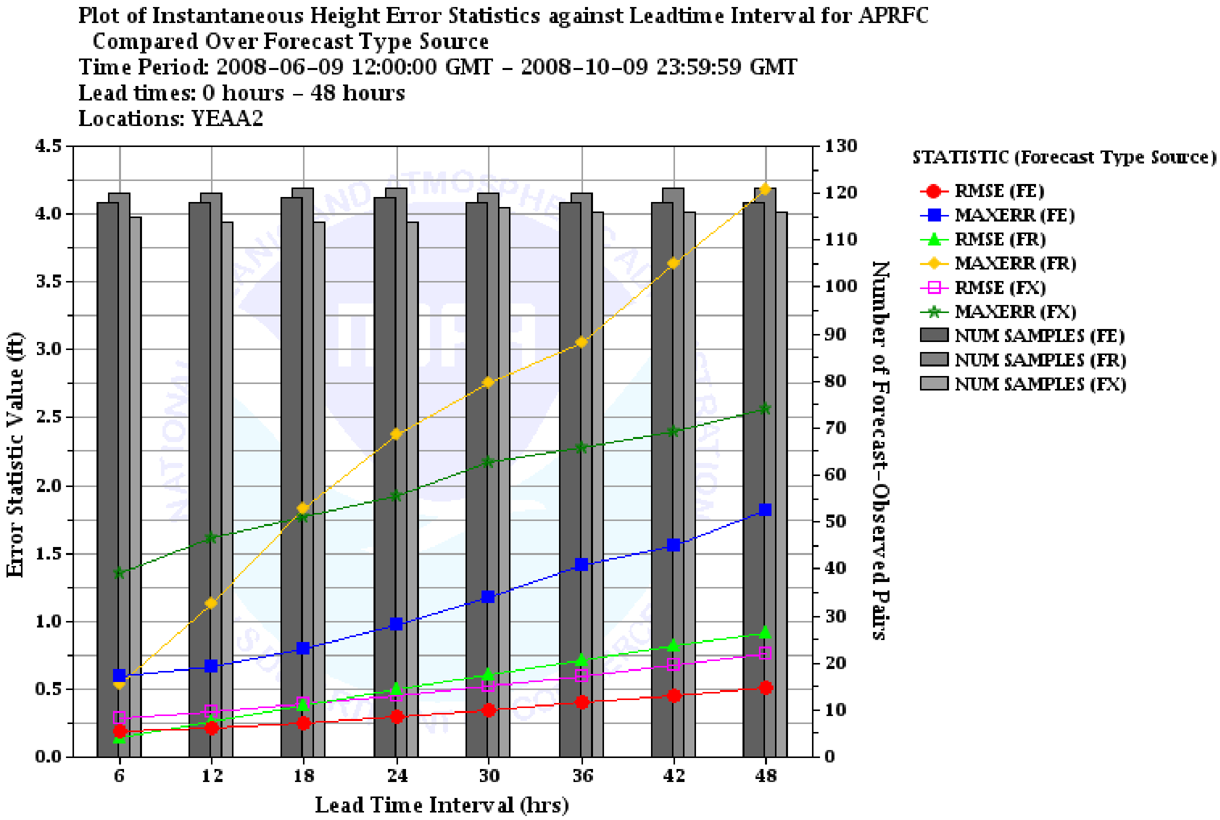
Below is a plot of Mean Absolute Error (MAE) and Mean Error (ME) for each of the forecast types (FE – new model forecast, FR – persistence, FX – old model forecast) by lead time.



Except for the first (6 hr) forecast, the new model had the lowest MAE for all forecast lead times and the improvement was greater with increasing lead time. The first value showing persistence with the lowest score is likely due to the fact that the persistence forecast was created in October for all forecast times and, therefore, likely included observed data either not available or not QC'd at the original forecast day/time. For the mean error (ME), both the new (FE) and old (FX) model forecasts showed an initial positive bias decreasing with forecast lead time and eventually turning into a negative bias at 30 hours for the FE's and at 42 hours for the FX's. The persistence ME remained near zero for all lead times. This makes sense when one realizes that there were near an equal number of forecasts for rising as for falling limbs of the hydrograph.

Future addition of QPF to the new model forecast should help remove the negative longer term biases. The RFC plans to implement QPF in these segments next open water season.

The next plot is of Root Mean Square Error (RMSE) and Maximum Error (MAXERR) for each forecast type by lead time.



Again except for the 6 hour lead time forecasts, the new model forecast (FE) showed less error in both categories with greater differences over longer lead times. The 6 hour persistence (FR) forecast was slightly better due to the same issues cited above.

In conclusion, we feel that the verification case study showed that calibrations based on reanalysis data were better than the old routed method used in the past. This has encouraged us to continue using these data for calibrations in historically data sparse areas in Alaska and several basins will be calibrated this development season using these data. We will continue to analyze forecasts for these and other basins using the verification tools provided by IVP and hope to expand our analyses by adding QPF to the model and comparing those forecasts with non QPF forecasts, the old model, and persistence. We also hope to be able to create forecasts for this and other basins using varying lead times of QPF (6, 12, 24, 48, 72, etc.) to determine the most appropriate amount of QPF to use for each forecast group or climatologic region.

WGRFC Verification Case Study
Presented on April 22, 2008 by Greg Waller

1. Introduction

Variational Assimilation (VAR) at West Gulf River Forecast Center (WGRFC) is a joint effort between Office of Hydrology (OH) and staff at WGRFC. D. J. Seo and Lee Cajina (OH), along with Tracy Howieson and Bob Corby (WGRFC) were instrumental in setting up VAR at WGRFC.

VAR basically “back calculates” soil moisture parameters using current observations as a starting point. VAR forecasts are only available at headwater points. Figure 1 shows the VAR basins across the WGRFC area. WGRFC has run VAR since 2003, and a sufficient data set exists to generate verification results. These results were compared to WGRFC operational forecasts.

2. Methods

VAR generates forecasts in 1 hour ordinates out to 72 hours. Synoptic hour runs (00Z, 06Z, 12Z, and 18Z) were archived. The operational model (National Weather Service River Forecast System (NWSRFS)) generates forecasts in 6 hour ordinates (usually) out 120 hours. Every forecast is archived. VAR forecast time series are easily displayed in the Site Specific Hydrologic (SSHP) model (primarily for WFOs) thanks to the work of Chip Gob at OH. Verification statistics were generated using the Interactive Verification Program (IVP). The work of Hank Herr, OH, in developing the IVP software was crucial in attempting this study.

The sheer amount of VAR forecasts cannot be trivialized. WGRFC runs VAR at 23 forecast points. 4 forecasts per day, with 72 hour ordinates, for 23 points, lead to 39,744 discreet forecasts daily from VAR output. This can tax the archive system when pairing data.

Forecasts used for this case study were restricted to the forecast points Sabine River at Greenville (GNVT2), East Fork Trinity River at McKinney (MCKT2), and Pine Island Bayou at Sour Lake (SOLT2) for March, 2008. SOLT2 is also a daily operational forecast point. GNVT2 and MCKT2 are flood forecast points.

3. Results and Discussion

Figures 2, 3, and 4 show paired data for the forecast sites GNVT2, MCKT2, and SOLT2 respectively. Figures 5, 6, and 7 indicate bias scores and sample size. More statistics were generated and can be made available on request. March, 2008, witnessed several flood events.

Operational forecasts were issued. Generally speaking, the operational forecasts were more than acceptable for WGRFC standards. These sites are generally fast responding and highly dependent on precipitation. The VAR forecasts (FA and FC columns) routinely "scored better" than the operational forecasts (FF column). Several factors lead to this conclusion.

4. Summary

VAR forecasts generally performed better than operational forecasts, based on statistics. However, a significant amount of the improvement of VAR over NWSRFS is the fact that VAR forecasts are issued more frequently, in smaller time steps, for a shorter duration. "Bad" forecasts can be absorbed in the amount of forecast data.

The primary factor driving the improvement of the VAR forecasts over the NWSRFS is the sheer volume of VAR forecast ordinates paired (72 ordinates 4 times per day (288 ordinates) vs. 20 ordinates 2 times per day (40 ordinates) with NWSRFS). One possible solution: restrict the VAR forecasts to the first 12 hours. This would allow for similar number of pairs (VAR – 48 vs. NWSRFS – 40). However there would still be the case of comparing apples to oranges: VAR out 12 hours as compared to NWSRFS out 120 hours. To derive meaningful statistics, more NWSRFS forecasts are needed. The best solution from a verification standpoint would be to make these forecast points "daily" forecasts. Over time, this would generate enough of a sample size for meaningful comparisons. There still would be the case of VAR producing significantly more forecast ordinates as NWSRFS. But a large enough sample size of data can still generate meaningful statistics.

Also, in an effort to standardize the verification results, we can try to restrict the data used to "synoptic" (6, 12, 18, 24, 30, 36, 42, 48, 54, 60, 66, and 72 hours) lead times from the 12Z forecast issuance only. This effort would greatly reduce the volume of VAR forecasts and would make them more comparable to the NWSRFS forecasts. The only caveat: NWSRFS forecasts would still need to be generated daily. Both solutions require more NWSRFS forecasts.

5. Figures

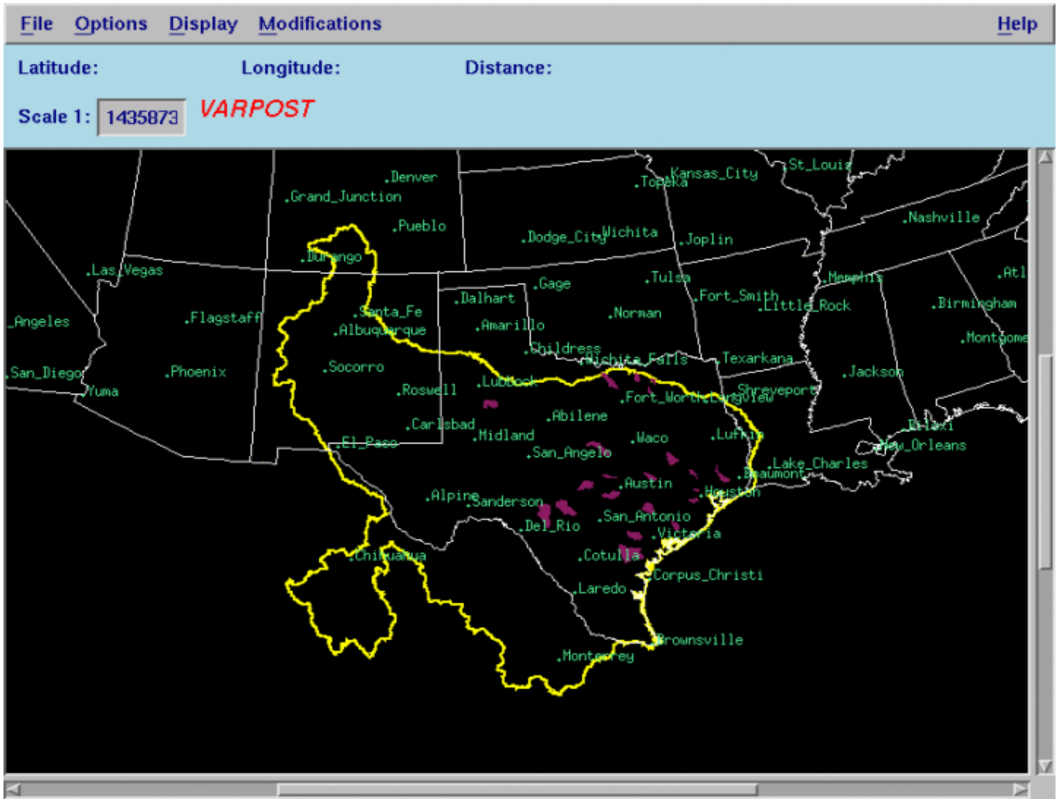


Figure 1. Area of interest.

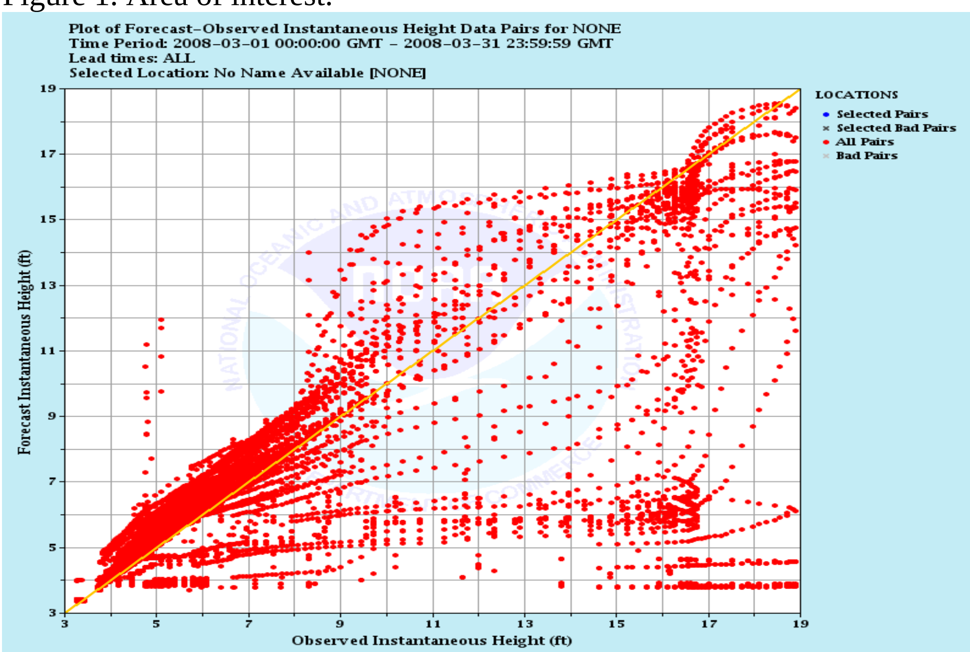


Figure 2. Paired data for GNVT2.

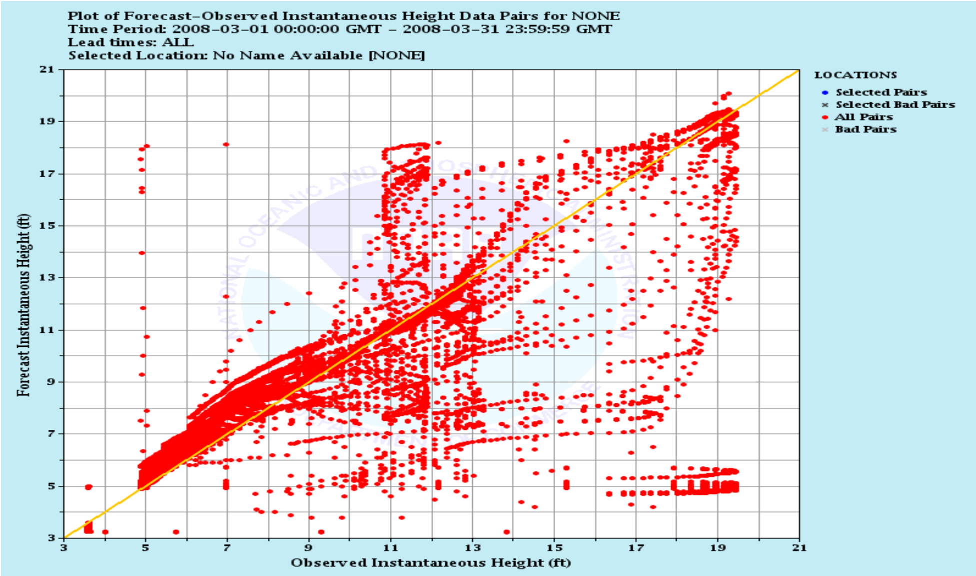


Figure 3. Paired data for MCKT2.

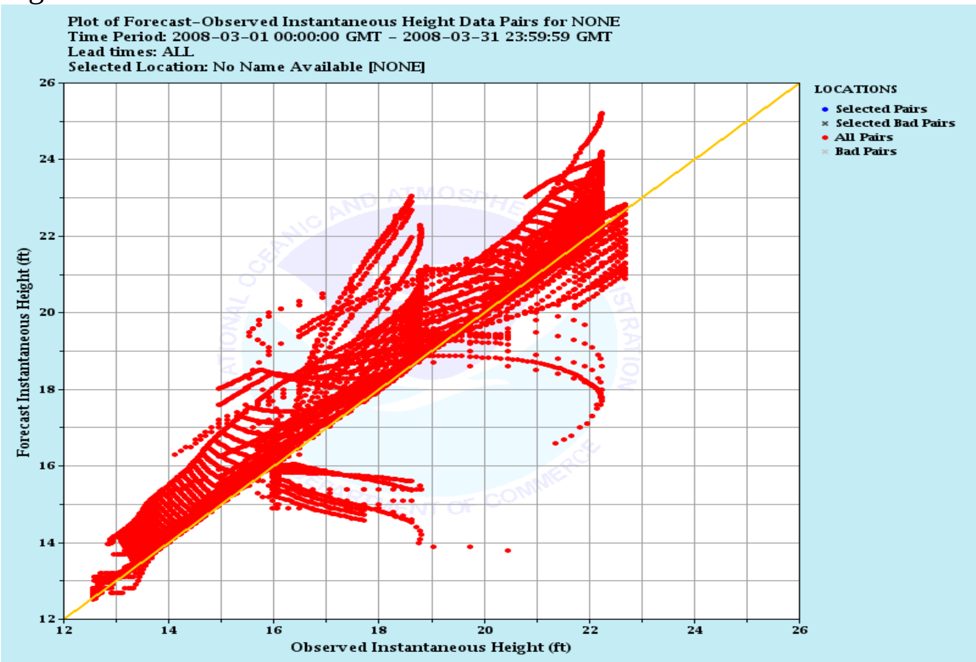


Figure 4. Paired data for SOLT2.

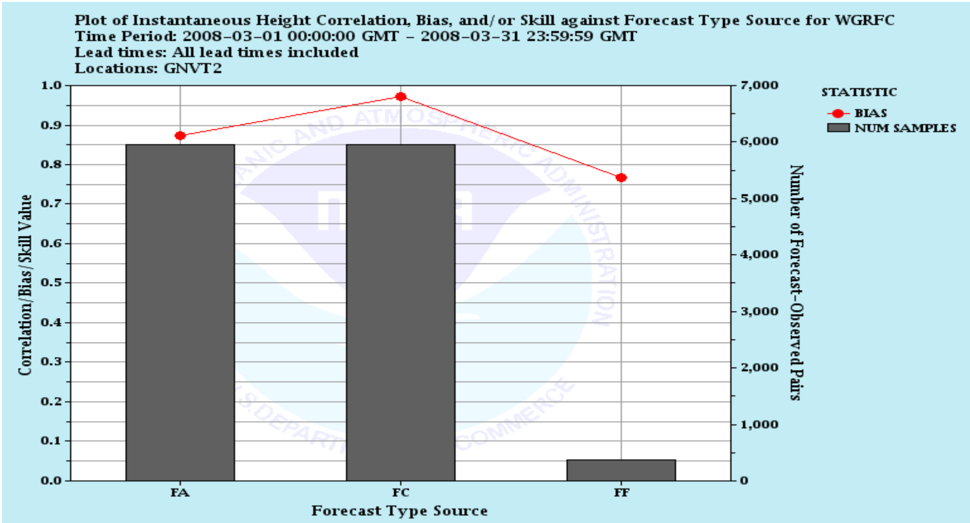


Figure 5. Bias and sample size for GNVT2.

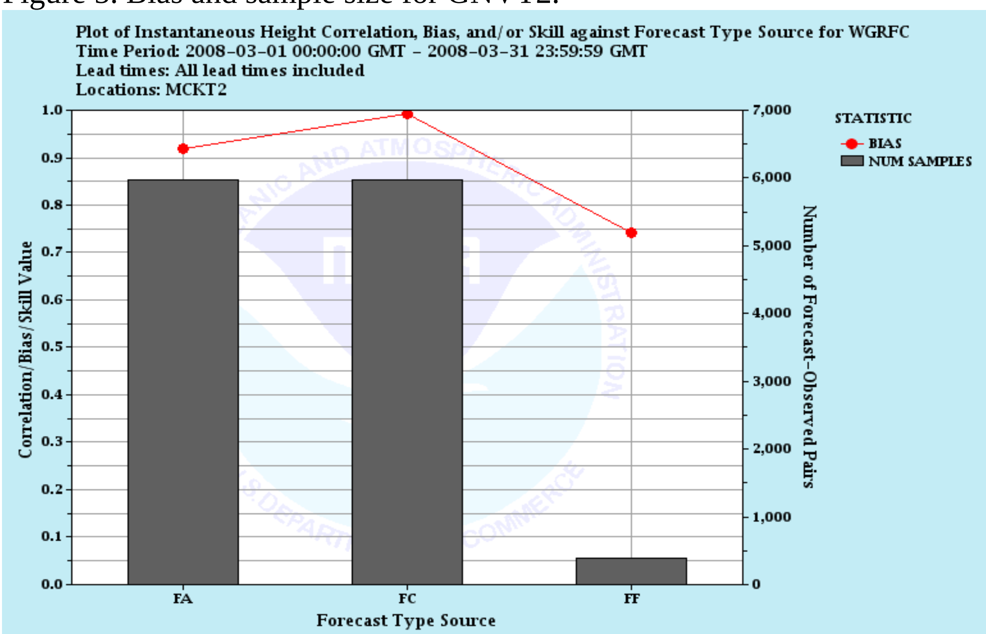


Figure 6. Bias and sample size for MCKT2.

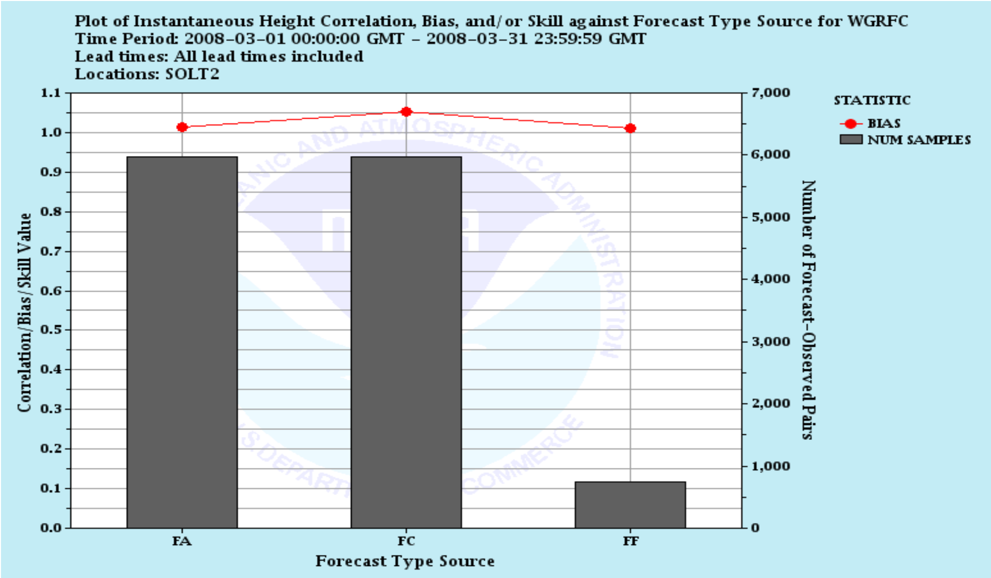


Figure 7. Bias and sample size for SOLT2.

For Figures 5 – 7, FA represents VAR forecasts with no human interaction, FC represents VAR forecasts with humans "balancing" the model parameters weekly, and FF represents the NWSRFC operational model.

ABRFC Verification Case Study
Presented on November 10, 2008 by Bill Lawrence

Summary

The ABRFC performed a verification study using EVS on 3 different basins using output from HMOS. The basic premise of the study was to determine if the HMOS procedure would yield reliable results while at the same time being able to distinguish between real events and false alarms. The ABRFC is eager to start producing short term probabilistic forecasts for its users and was hoping HMOS would prove useful in this manner. Earlier attempts to produce useable short-term probabilistic forecasts have thus far been unsuccessful as the end product verified as unreliable. It has been determined that earlier attempts did a good job including the meteorological uncertainty associated with river forecasting, but did not include any hydrologic uncertainty which lead to poor overall results. It is hoped that HMOS will be able to produce forecasts that have both meteorological and hydrologic uncertainties included, leading to an overall reliable short term probabilistic forecast, something that will be very helpful to our customer base.

Data

HMOS requires a long history of deterministic forecasts. The ABRFC is lucky to have a long archive of operational forecasts stored in the Postgres relational database that have a 6 hour time-step and produce forecasts out to 120 hours. . However, only our daily forecasts will likely have a sufficient number of forecasts from all types of situations (low, medium and high flows) to be able to produce HMOS statistics for. For this study, 3 of our daily forecast points were used. These points include BLUO2 (Blue River near Blue, OK), QUAO2 (Spring River near Quapaw, OK), and DEKT2 (Red River near Dekalb, TX). BLUO2 is a headwater point on a difficult to forecast and well studied basin. It is the smallest of the 3 basins, but has a very long and narrow basin. Both QUAO2 and DEKT2 are not headwaters. QUAO2 is a moderately sized basin, with several basins located upstream. DEKT2, on the other hand, has a large basin above it contributing to its flow. DEKT2 is also heavy regulated as it is downstream of Denison Dam. Denison Dam is a power generating dam, thus flows are sometimes regulated for the economic benefit of power generation, especially during the warmer months of the year.

11 years worth of deterministic forecasts were used for this study, from Feb 5, 1997 until Oct 31, 2008.

Method

HMOS was run in a hindcast mode using the 11 years worth of deterministic forecasts stored in the Postgres database. All forecasts issued were included in the study. During times of flooding, additional forecasts are issued in addition to the daily 12z forecast. In some cases, more than one forecast can be issued for the additional synoptic forecast periods (18z, 00z, 06z), especially in rapidly changing flood conditions. This ultimately yields more than 365 forecasts for each point for each year.   The Postgres database was also used to provide a “current” stage used by HMOS for each deterministic forecast.

It should be noted that the hindcast procedure used for HMOS is very effective, and easy to use. The run length for each HMOS forecast found in the archive is approximately 25-40 seconds, yielding several hours to run each year’s worth of data. An 11 year period of hindcasts was normally run by command line, and let run overnight until the entire period was completed.

The EVS software was then used to produce verification metrics from the resulting .CS files created by HMOS.   EVS had no problems reading in the thousands of forecasts and producing results. Several suggestions were noted for EVS improvement and passed along to OHD.

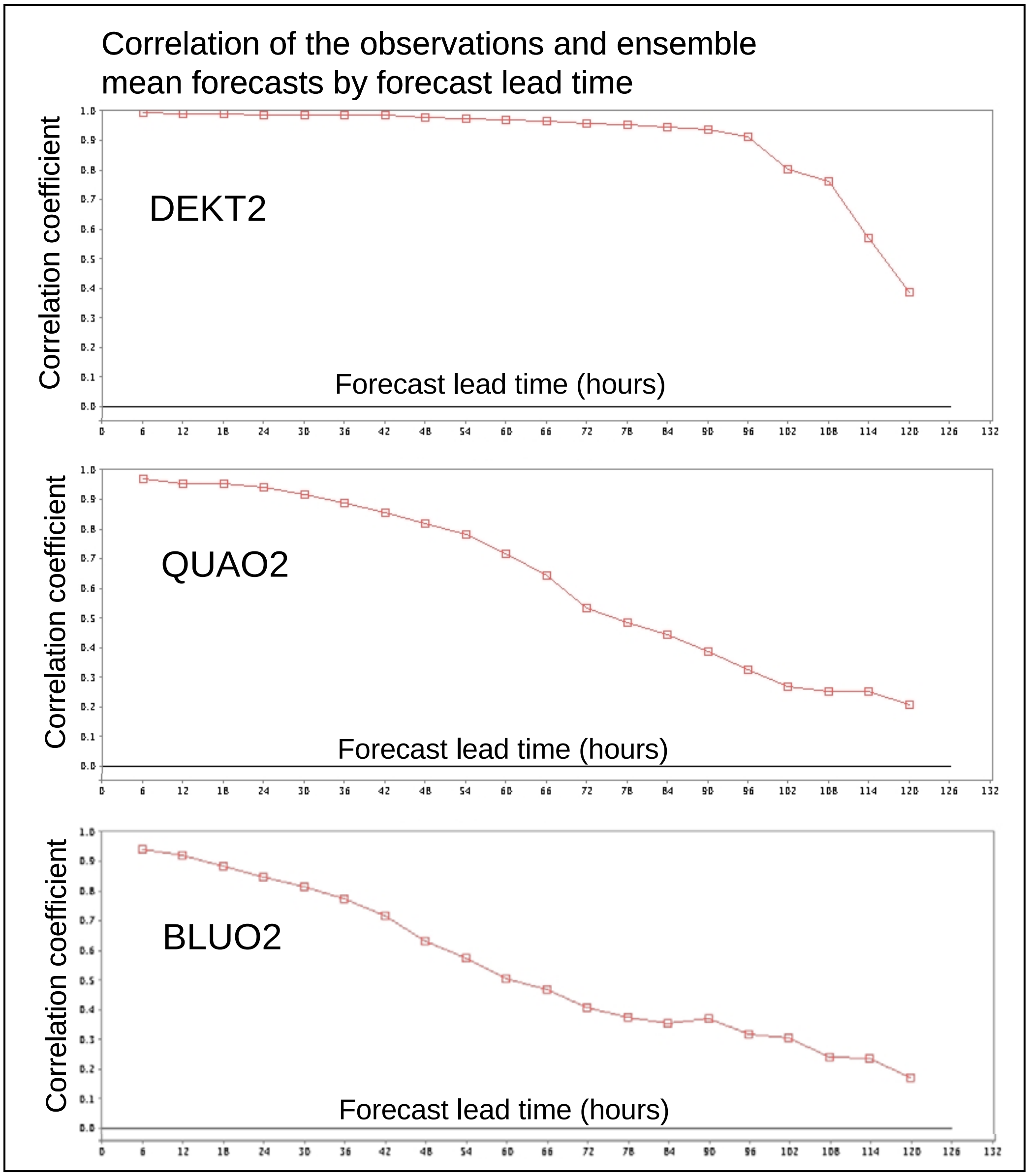
Results

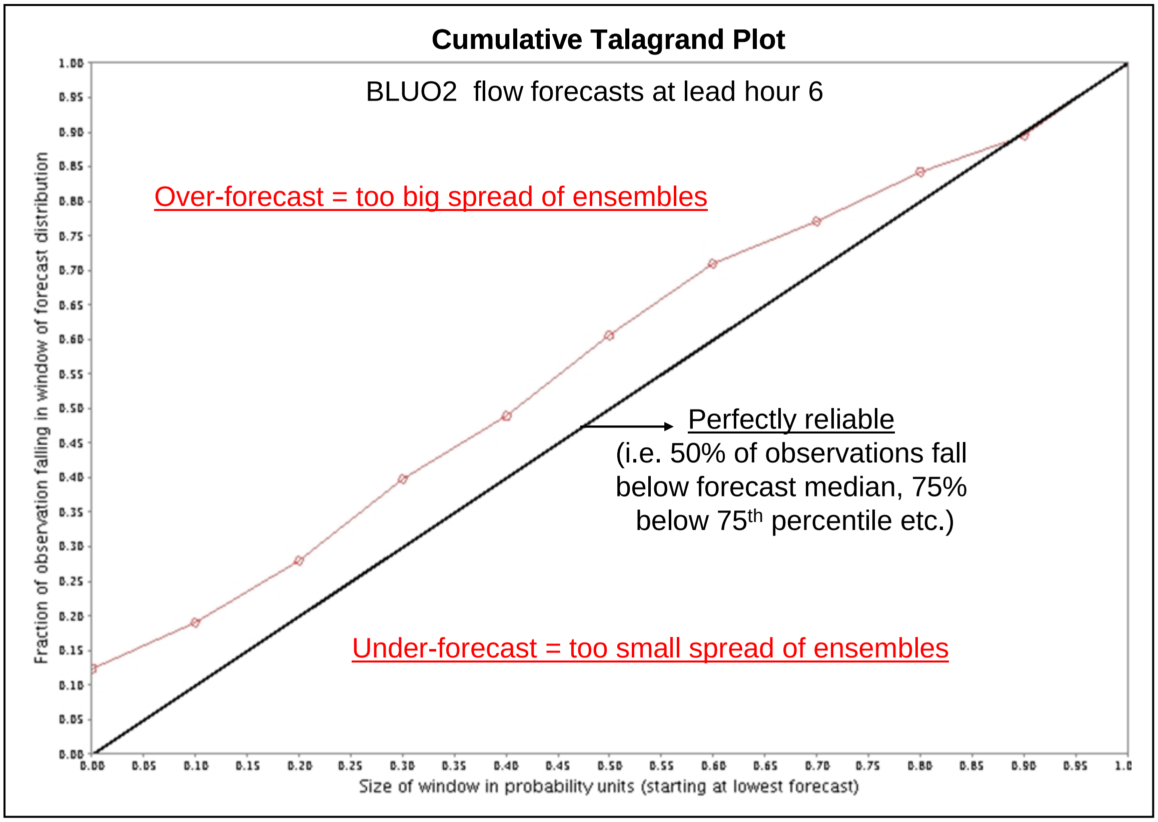
Overall, the results proved quite positive. All 3 points showed results that were reliable. The Cumulative Talagrand diagrams were used mostly to determine reliability (see graphics below). As would be expected, the reliability of the HMOS forecasts decreased slightly as lead time increased. However, these results show that for the most part, HMOS forecasts are reliable. This is especially the case when comparing results of HMOS to the prior method of short-term probabilistic forecasts used at the ABRFC.   There has been some concern that HMOS may fail to show extreme events, when heavy QPF just outside of the ABRFC’s 12 hour window for use in the deterministic forecast will cause significant rises. However, a recent look at the data over the years showed that for all 3 points, a miss of greater than 6 feet in stage forecast only occurred less than 2% of the time.   This should fit within the statistical bounds given by HMOS. Another concern is the use of only one rating curve (the most current one) by the ABRFC to convert the archived stage forecast data as well as observations into flow, which is used by HMOS.   Future study or evolution of HMOS should consider using archived ratings for the entire 11 year period and see if results improve. Also, it may be worth downloading raw flow data from the USGS for observations and see if that affects the output metrics much.

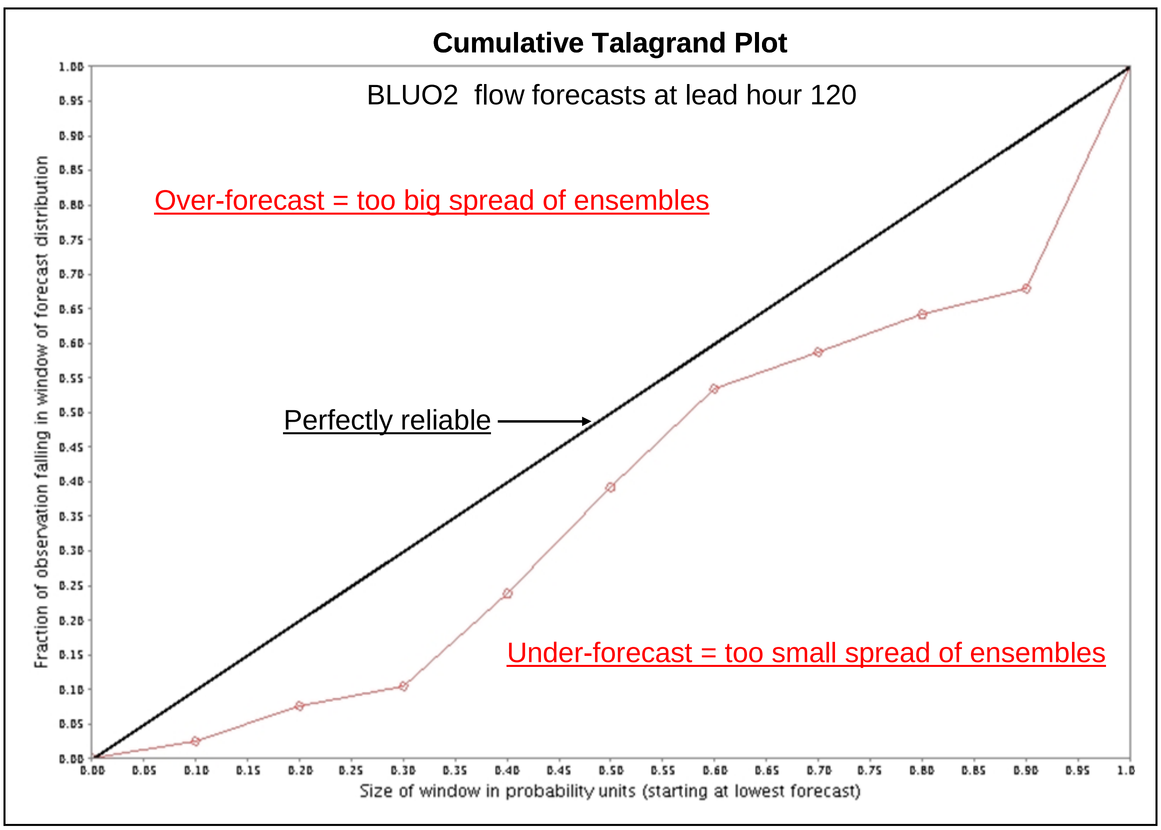
The correlation coefficient of the HMOS mean forecast was highest for DEKT2, the largest total basin of the 3 points studied (see graphics below). This makes a lot of sense as surprise local rainfalls would tend to have a much more muted appearance in the hydrograph as compared to the smaller basins of BLUO2 and QUAO2.   Overall, correlation between the mean of the HMOS ensembles decreased as the basin size decreased.

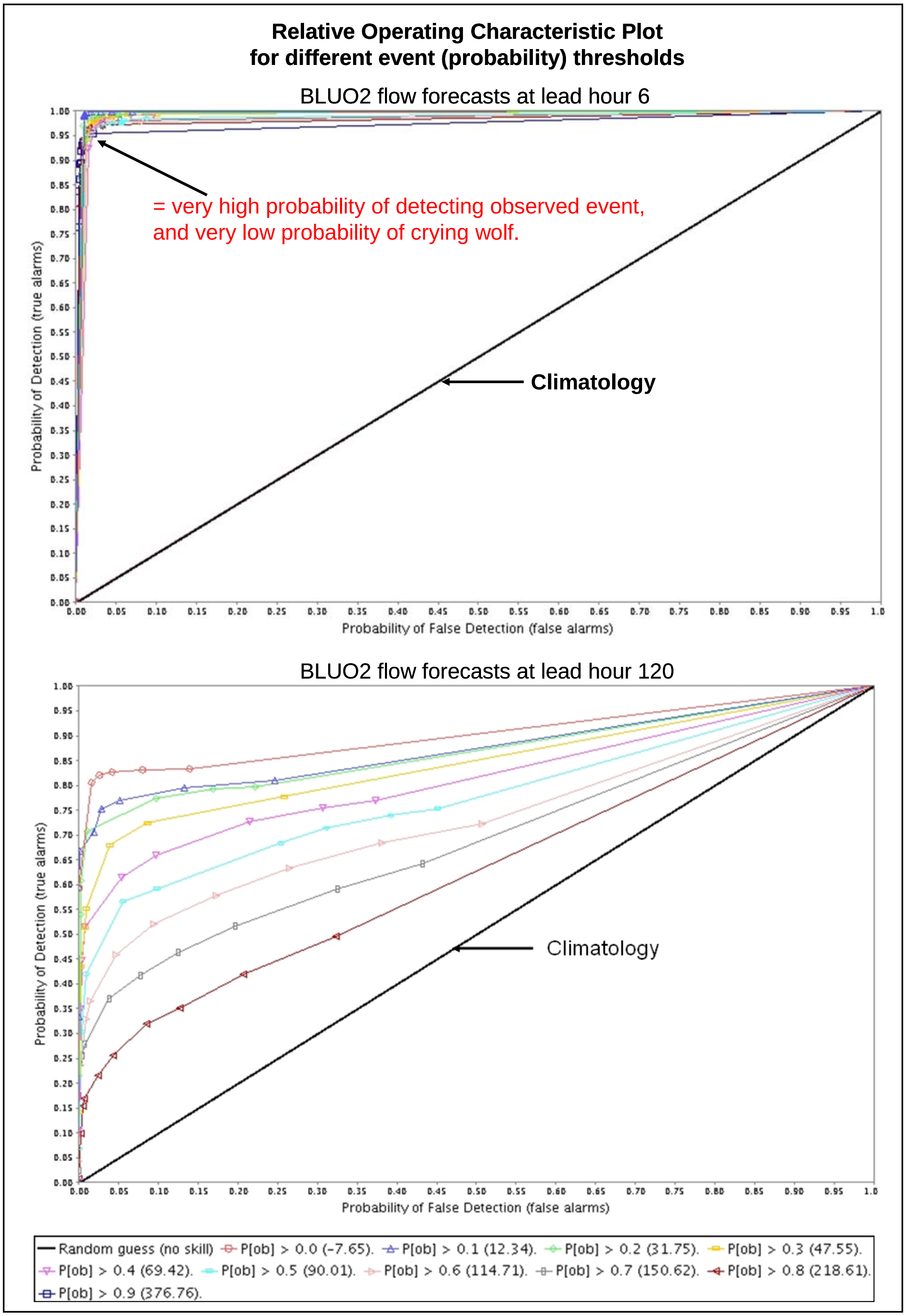
The Relative Operating Characteristics (ROC) charts obtained from EVS showed that HMOS has a good ability to discriminate between events. As with the reliability, the skill decreases with increasing lead time.   But the results were deemed acceptable.

After reviewing all the EVS produced metrics for all 3 basins, it was also became clear that overall, results were the best for the largest basin, and were worst, relatively speaking, for the smallest basin.


Correlation of the observations and ensemble mean forecasts by forecast lead time
DEKT2
QUAO2
BLUO2
Correlation coefficient
Forecast lead time (hours)


Cumulative Talagrand Plot
BLUO2 flow forecasts at lead hour 6
Over-forecast = too big spread of ensembles
Perfectly reliable
(i.e. 50% of observations fall below forecast median, 75% below 75th percentile etc.)
Under-forecast = too small spread of ensembles
Fraction of observation falling in window of forecast distribution
Size of window in probability units (starting at lowest forecast)


Cumulative Talagrand Plot
BLUO2 flow forecasts at lead hour 120
Over-forecast = too big spread of ensembles
Perfectly reliable
Under-forecast = too small spread of ensembles
Fraction of observation falling in window of forecast distribution
Size of window in probability units (starting at lowest forecast)


Relative Operating Characteristic Plot
for different event (probability) thresholds
BLUO2 flow forecasts at lead hour 6
= very high probability of detecting observed event,
and very low probability of crying wolf.
Climatology
Probability of Detection (true alarms)
Probability of False Detection (false alarms)
BLUO2 flow forecasts at lead hour 120
Climatology
Probability of Detection (true alarms)
Probability of False Detection (false alarms)
Random guess (no skill)
P[ob] > 0.0 (-7.65).
P[ob] > 0.1 (12.34).
P[ob] > 0.2 (31.75).
P[ob] > 0.3 (47.55).
P[ob] > 0.4 (69.42).
P[ob] > 0.5 (90.01).
P[ob] > 0.6 (114.71).
P[ob] > 0.7 (150.62).
P[ob] > 0.8 (218.61).
P[ob] > 0.9 (376.76).

Unmet needs with current EVS prototype software

- Need for EVS to be able to create various types of box plots from the GUI interface (not by modifying the *.evs project file).
- Need for EVS to be able to create sample size plots for various metrics.
- Would like the EVS' User Manual to contain easy to understand and real examples of graphics from EVS, as well as simple to understand explanations of results (Talagrand graphic that is doctored up with labels). It would be nice to have these labels inside the application as well (instead of links or mathematical formulas)
- Need more user friendly error messages, to better explain what went wrong.
- Need more intelligent way to read in .CS files (not just the entire directory).

MARFC Verification Case Study
Presented on November 10, 2008 by Andrew Philpott

In this case study, the MARFC verified short term Ensemble Streamflow Prediction (ESP) forecasts for the Juniata River using the Ensemble Verification System (EVS). Due to data limitations, the period of record was February 2006 through June 2008. This is not an ideal length for ensemble verification, but is a good starting point for beginning to understand the characteristics of the errors in these forecasts. Since EVS is able to compute statistics for any ensemble forecast, MARFC also verified the Ensemble Pre-Processor (EPP2) precipitation and temperature traces which serve as the forcings for the ESP forecasts. MARFC looked at statistics for most of the forecast points within the Juniata basin, but our case study analysis concentrated particularly on a headwater point (Spruce Creek on the Little Juniata River) and a downstream point (Newport on the Juniata River). See Figure 1 for a map of the basin.

The MARFC short term ensemble streamflow forecasts are being prepared in collaboration with the Hydrologic Ensemble Prediction group at OHD/HSMB. These forecasts go out to a lead time of 7 days. MARFC produces the forecasts once per day at 12z. The EPP2 preprocessor has been calibrated to account for errors in our deterministic QPF and temperature forecasts. There are 50 (fifty) 7-day future rainfall and temperature scenarios which are used as forecast inputs for the Continuous API rainfall/runoff model to create 50 ESP traces. The first 2 days worth of probabilistic quantitative precipitation forecast (PQPF) are based on the official MARFC deterministic QPF, while the final 5 days are based on smoothed climatology. For temperature, the first 5 days are based on MARFC deterministic forecasts, whereas the final two days are based on climatology. The initial model hydrologic state and operational modifications are the same in each of the 50 runs.

An example of an ESP forecast for Spruce Creek is given in Figures 2 and 3. Figure 2 shows the PQPF traces. The time step is 6 hours. During the first 18 hours, the operational QPF was relatively high, which lead to a large spread among the PQPF members. Afterwards, the operational QPF decreased substantially, eventually reaching zero. At a QPF of zero, nearly all of the members were also zero, and the total spread was less than 0.10 inch. Finally, for the last five days the climatological PQPF had a spread around 0.50 inches. If the pure historical record were being used, rather than a smoothed climatology, there would be some values higher than 0.50 inches. Figure 3 shows the expected value plot for the streamflow forecast. During the first 6 hours, there was very little spread between members, since all traces began from the same initial hydrologic state. Since Spruce Creek is a rapidly responding headwater point, cresting within 6 hours of the rainfall, the QPF quickly produced a large spread by lead hour 12. This spread gradually decreased as the low QPF and climatology based PQPF numbers began to dominate the total runoff.

In these streamflow forecasts, we identified three phases to the forecast. Forecasting error in each phase has distinct properties, but the transition between phases is continuous rather than abrupt. The first phase is dominated by the initial hydrologic condition, including simulated runoff from observed precipitation. Because the forecasts lack any hydrologic uncertainty or uncertainty in the initial conditions, these forecasts are severely under-spread. In the second phase, the QPF based PQPF generates most of the simulated runoff. In this phase, spread has increased due to spread among the QPF values, but there is still no accounting for hydrologic uncertainty, and therefore the forecasts are still under-spread. In the third phase, most of the simulated runoff is from the climatology based PQPF. Discrimination between events and non-events decreases substantially, and under-forecasting bias for higher streamflow events increases. However, the model still retains memory of the QPF based PQPF and carryover state, so the forecasts have some skill in discriminating between events and non-events, which a pure climatological forecast will not have.

In phases two and three, some of the streamflow error quality can be linked to the errors in the PQPF forecasts. Unlike streamflow, PQPF has two very clear phases rather than a transition between phases. The QPF based PQPF from the first 48 hours is abruptly different from the climatology based PQPF. An example of this can be seen in the mean error in the ensemble mean (Figure 4), where the negative bias for higher rainfall events more than doubles when changing from QPF based to climatology based PQPF. Correspondingly, for high flow events, the streamflow forecast becomes more negatively biased with lead time, but in a more continuous fashion. One reason there is so much bias in climatology based PQPF is that the climatology is smoothed, and so no high events (> 1” in a day) will be in any of the climatological traces.

During phase one of the streamflow forecasts, which is only the first 6 hours at Spruce Creek, there is very little spread and hence low reliability due to severe under-spread (Figure 5). However, because the error magnitude is generally low, the discrimination between events and non-events is high, both when the event is defined as streamflow over 15 cms and 30 cms (Figure 6).

In phase two, the spread improves due to incorporation of QPF, particularly on the upper end of forecasts (Figure 7). Whereas in phase one, only 8% of observations were within the spread of forecasts from the pair that went with each of them, in phase two that has increased to 45%. The biggest problem with reliability is seen on the low end of the forecasts, when the base flow has a tendency not to recede quickly enough. Therefore, more than half the time the observation is below all of the 50 forecast traces. On the high end, there is much less under-spreading. Discrimination remains high (not shown).

In phase three, there is not much change in reliability, although an increase in spread for low flow forecasts increases the percentage of time that the observations fall in the forecast spreads to 55%. The biggest changes are the increase in negative bias for high flow events and the decrease in discrimination due to the fact that runoff is being generated by climatology based PQPF that have no discrimination skill (Figure 8). However, memory of observed rainfall and QPF based PQPF keeps some discrimination skill in the forecasts, meaning that there is a higher probability of detection than probability of false alarm. Overall, despite a slight increase in reliability for low flow forecasts from phase two to phase three, the phase two forecasts are superior to phase three, particularly for higher flow events. This is confirmed by several verification statistics, such as bias in the ensemble mean bias and continuous ranked probability score (MCRPS). MCRPS is a measure of the degree to which the probabilistic forecasts match up with the overall probabilities of various observed values.

We have further observed that the behavior of the errors is highly dependent on the response time of the basin to rainfall. For example, since Newport responds to a much larger basin than Spruce Creek (3354 square miles versus 220 square miles), it takes two and one half days longer to crest. Because of this, a statistic for Newport should be compared to a statistic for Spruce Creek from an earlier lead time. For example, the ROC curve at Newport at Lead Hour 90 is very similar to that for Spruce Creek at Lead Hour 30, but the curve at Lead Hour 30 is quite different (not shown). This shows that basins of a different size and response time should not be aggregated to increase sample size, as the combined dataset will not be interpretable. Also because of the three phases of the streamflow forecasts, caution should also be used in aggregating temporally (such as a daily average).

The temperature forecasts verified well for the first five days, with a mild under-forecasting bias, generally good reliability, and high discrimination (not shown).

Overall, despite a limited dataset, MARFC has discovered, through EVS software, several properties of the error in our short term ESP forecasts. QPF based rainfall forecasts in the first two days are superior to the climatology based rainfall forecasts from days three through seven. In particular, there is less bias in the forecasting of high events, and there is skill in discriminating between events and non-events. In streamflow forecasts, there is substantial error due to a lack of hydrologic uncertainty. This is particularly evident in the early lead times, where there is no spread between the traces since QPF has not been applied yet. However, under-spreading in the forecasts continues throughout the 7 day forecast, particularly for the lowest forecasts (for which QPF was low) where much of the under-spread is due to lack of simulated uncertainty in base flow recession. Using an ensemble post-processor to spread out the forecasts could improve the forecasts by artificially accounting for hydrologic uncertainty. For higher flow events, there is an under-forecasting bias in these forecasts beyond the first 6 hours due to under-forecasting bias in precipitation. As climatology based precipitation begins to dominate runoff generation, the under-forecasting bias increases substantially. Also, discrimination skill over climatology decreases as unskilled climatological QPF takes over runoff generation. Hence the earlier lead time forecasts based on QPF are of better quality than longer lead time forecasts based on climatology; using QPF adds value over climatology.

We have confirmed that there are a number of identifiable error sources in these forecasts, particularly under-forecasting of heavy rainfall events and lack of hydrologic uncertainty. These identify two areas of potential improvement. However, the forecasts do provide valuable information to the user. In particular, these forecasts have a good ability to discriminate between conditions that are likely to produce high streamflow events and conditions that are unlikely to cause significant streamflow elevation.

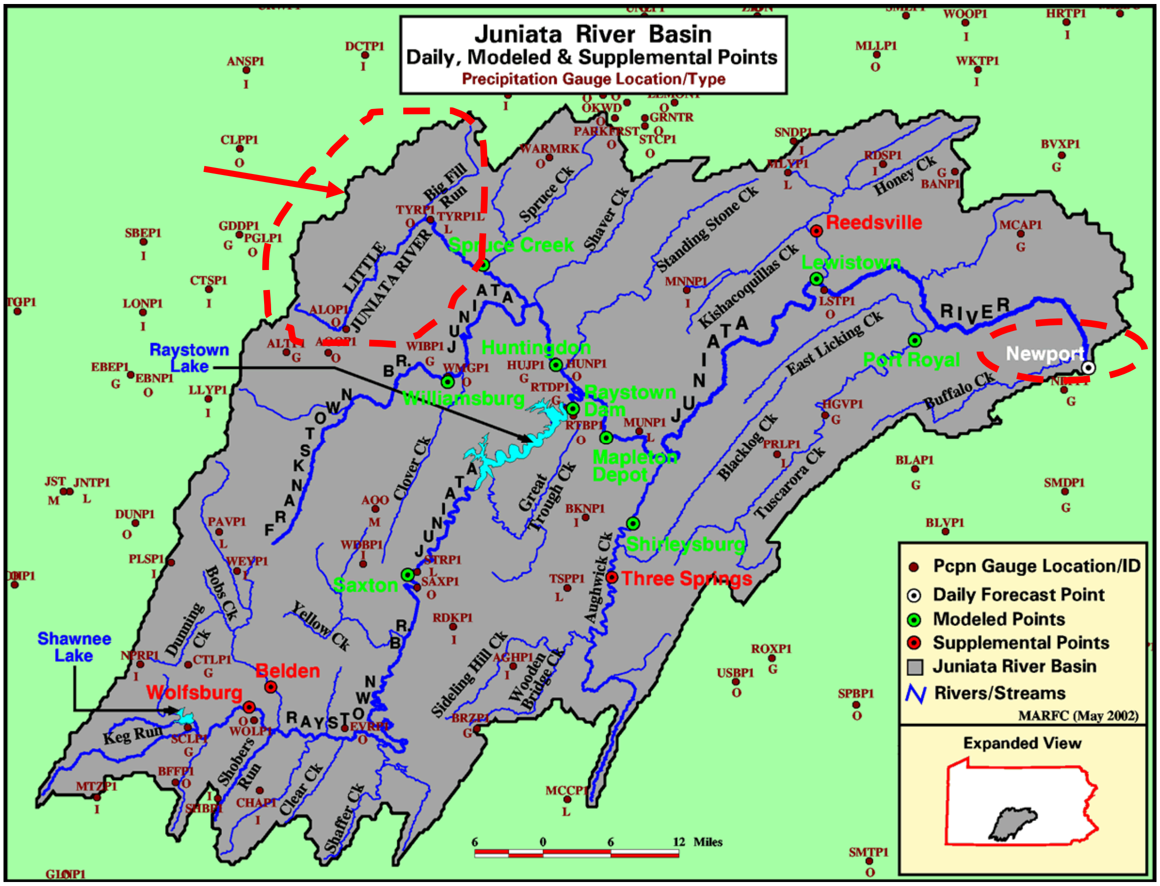
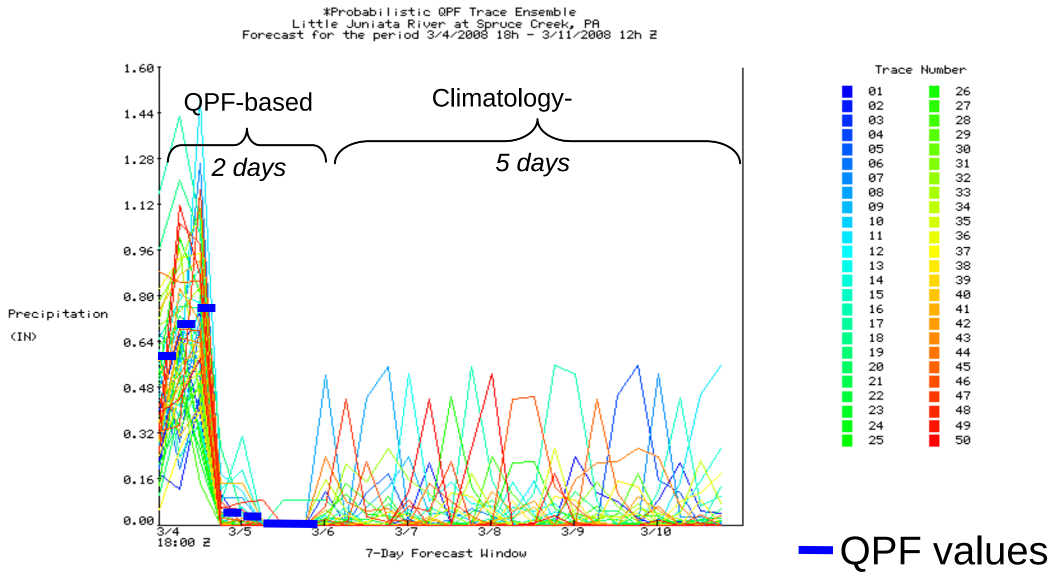


Figure 1. Map of the Juniata Basin with the Spruce Creek local drainage basin (Little Juniata River) and Newport gage highlighted. Generally the streamflow is southwest to northeast within this basin, with Newport as the downstream outlet. Map made by David Solano, Senior HAS for MARFC operations.



Deterministic QPF (for lead hours 6-48): 0.55, 0.72, 0.76, 0.04, 0.03, 0.00, 0.00, 0.00 in

Figure 2. Probabilistic QPF traces as generated by the EPP2 preprocessor for the short term ensemble forecast that was issued March 4, 2008 for Spruce Creek. Time step is six hourly, and total lead time is seven days. The first two days are based on the MARFC deterministic QPF, while the final five days are based on smoothed climatology.

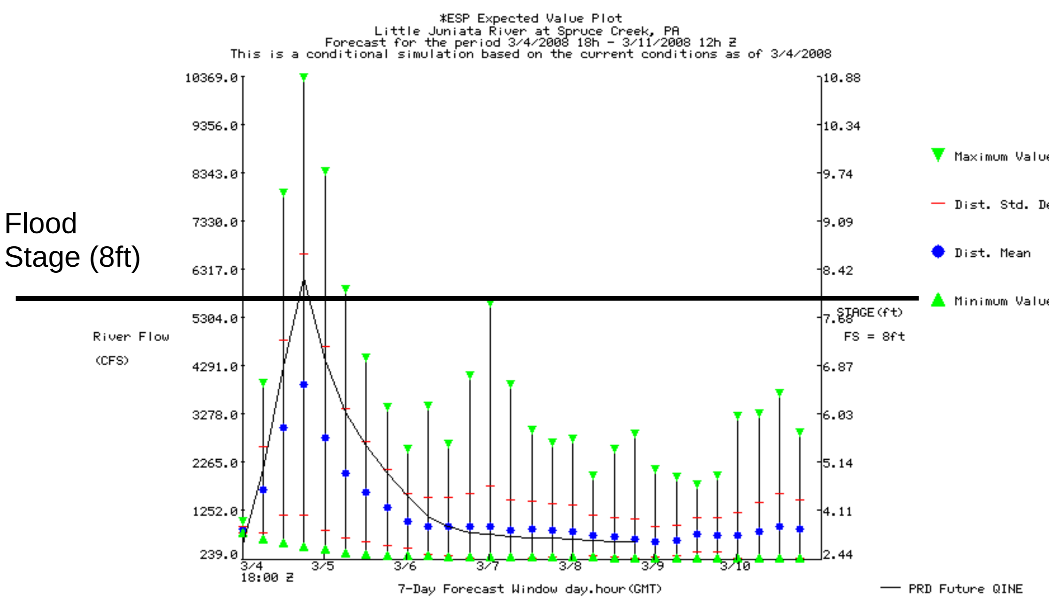
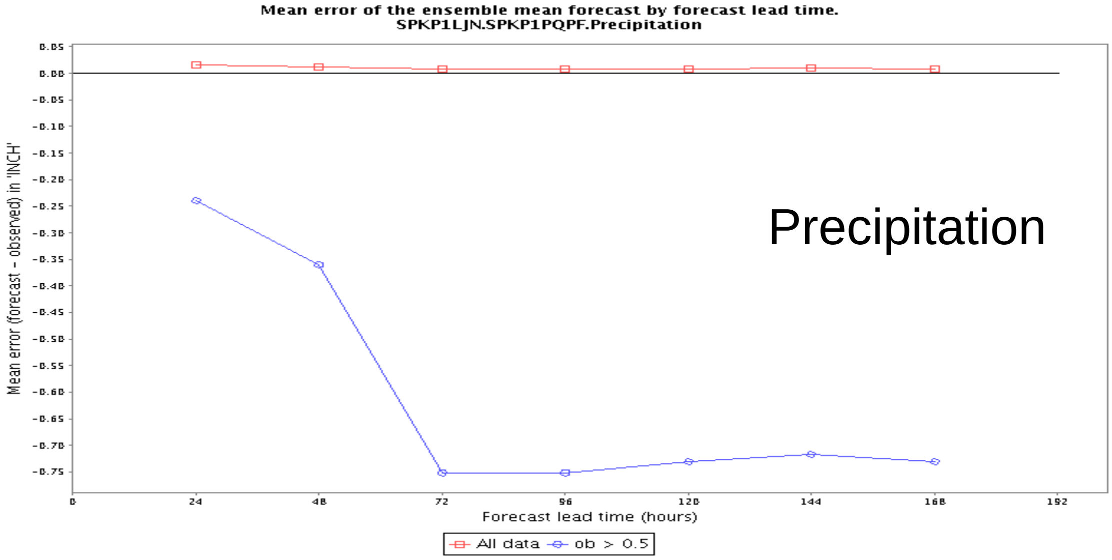


Figure 3. This ensemble streamflow forecast was issued March 4, 2008 from the traces from Figure 2. For each 6-hourly period, the highest streamflow forecast, lowest streamflow forecast, ensemble mean, and ensemble standard deviation range are shown.



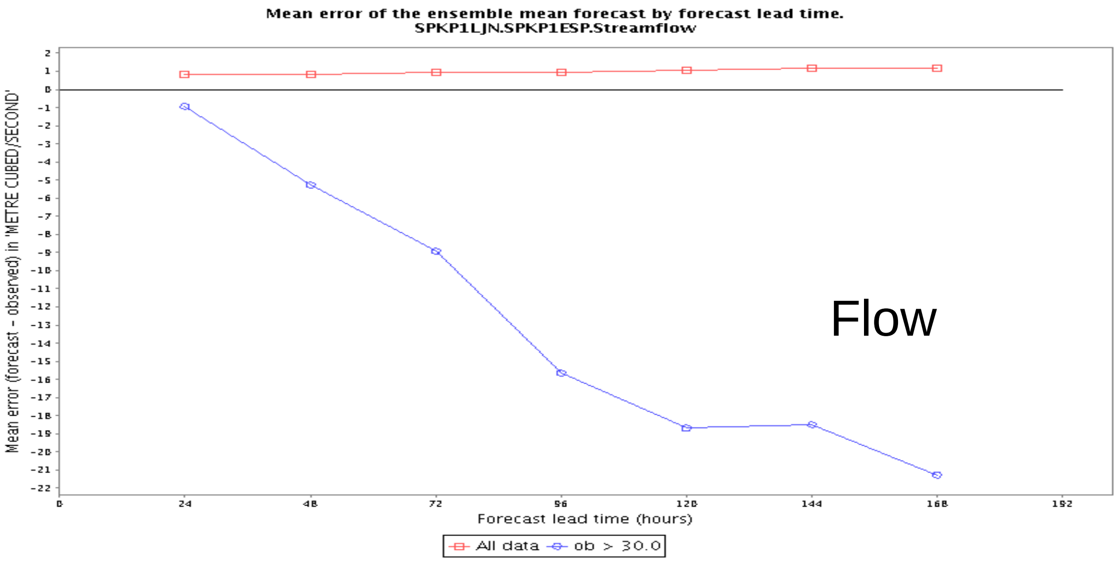


Figure 4. For all data, there is little bias in the daily total precipitation ensemble mean and daily average streamflow ensemble mean forecast. However, when only observation and forecast pairs are compared when there is rainfall above 0.5” or streamflow above 30 cubic meters per second (cms), there is a negative bias. For precipitation, the bias is twice as high using climatology (days three through seven) than using MARFC QPF. For streamflow, bias increases steadily with lead time, and is highest when being forced by climatology based PQPF.

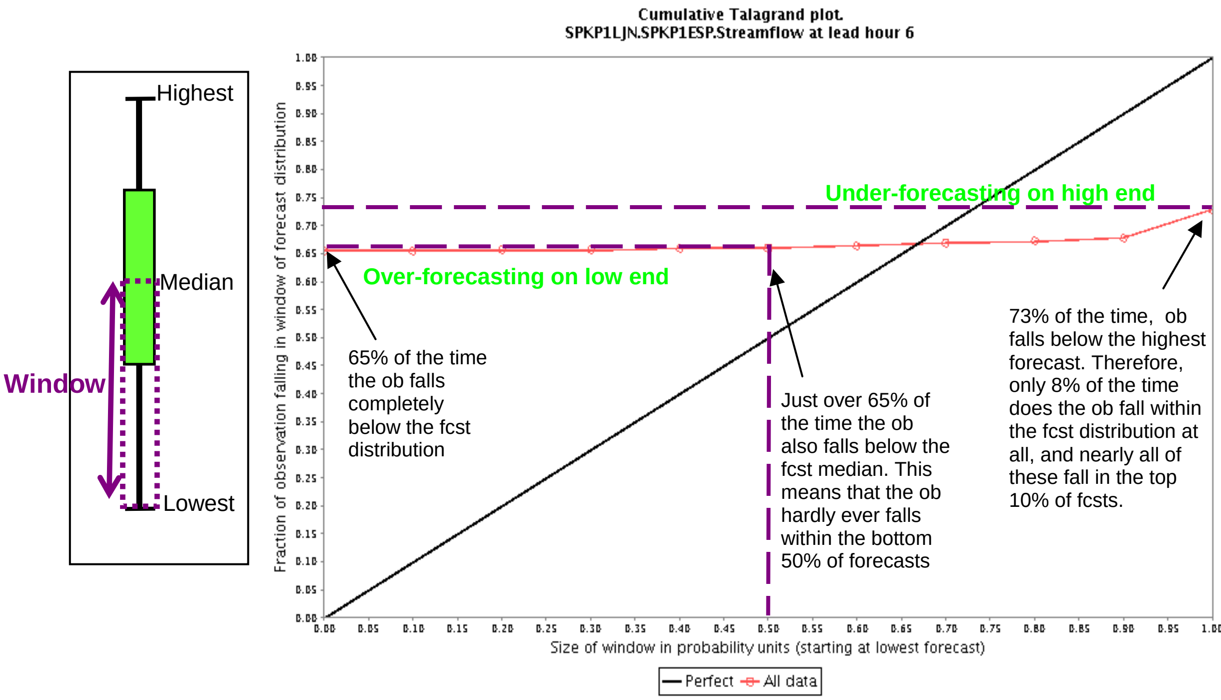


Figure 5. At lead hour 6, there is very little spread due to QPF. Because of a lack of hydrologic uncertainty the forecasts are severely under-spread, with the forecast falling within the range of 50 traces that pair with it only 8% of the time.

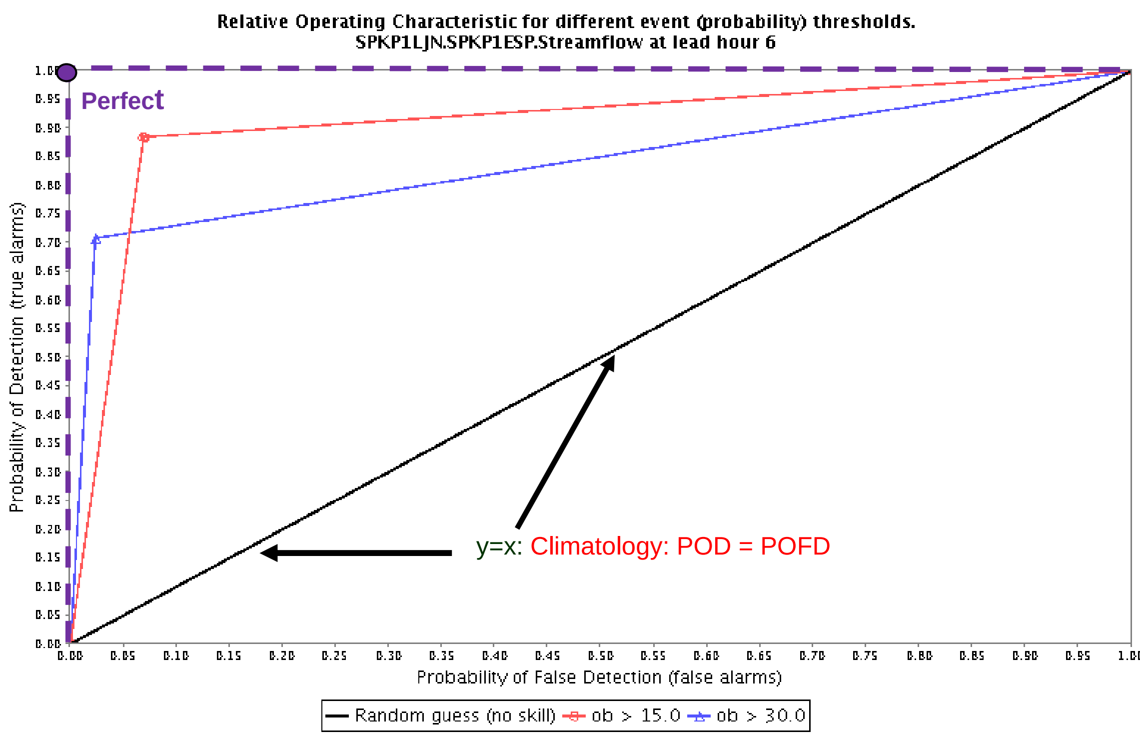


Figure 6. At lead hour 6, probability of detection (hits divided by hits plus misses) is substantially greater than probability of false detection (false alarms divided by false alarms plus true negatives). This means that discrimination skill is high.

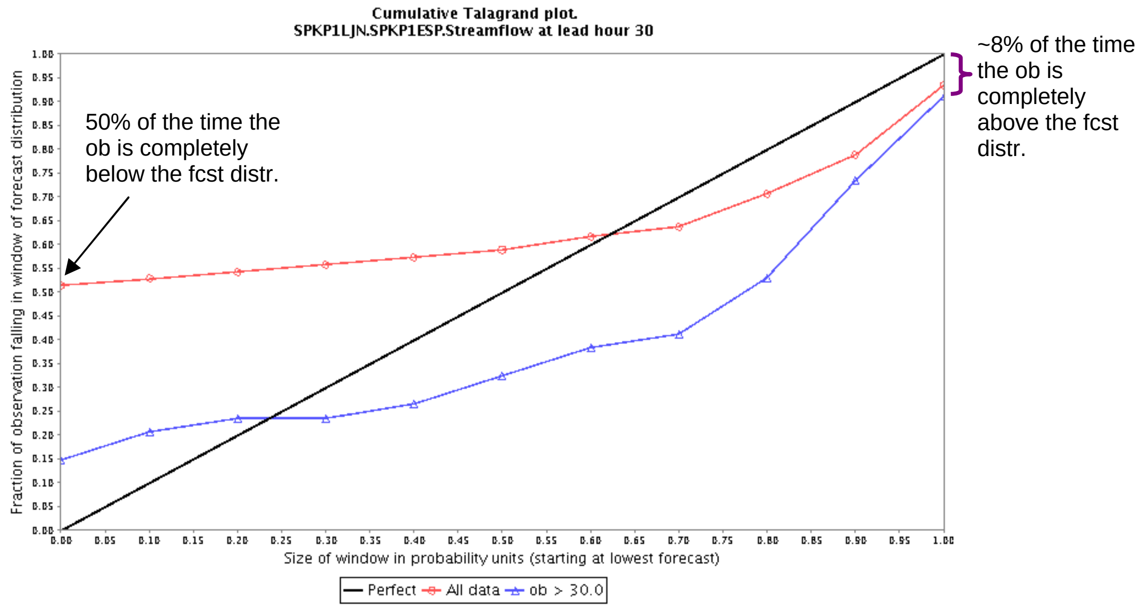


Figure 7. During phase two, when runoff is being generated by MARFC QPF-based PQPF, the spread is improved over the six hour forecast, with now 42% of the time the observation falling within the 50 forecast traces that pair with it. The biggest remaining problem in under-spread is on the lower end of forecasts.

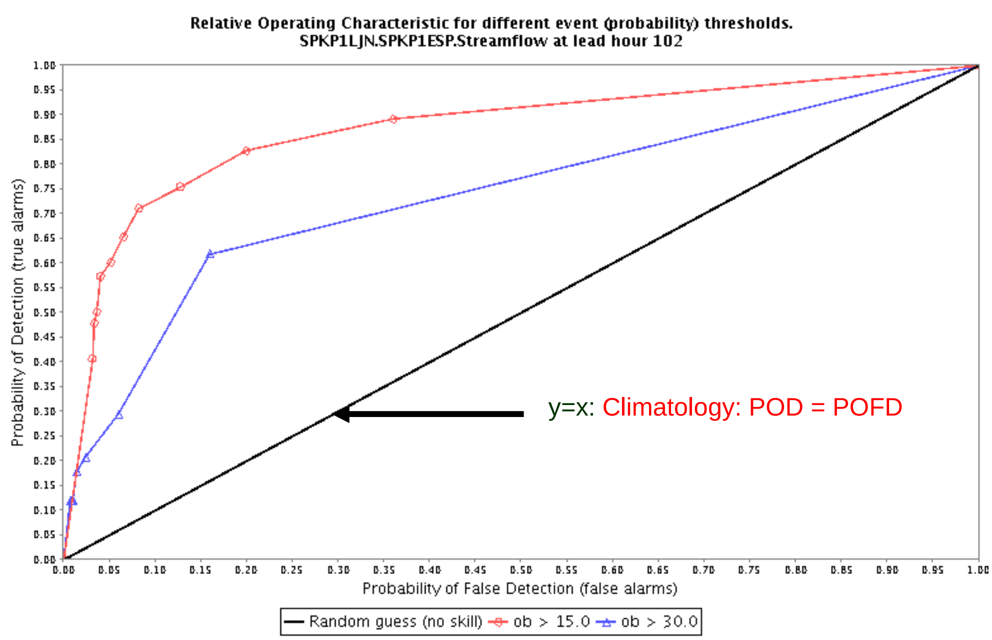


Figure 8. During phase three, when runoff is being generated by climatology-based PQPF, the discrimination skill is lower than for earlier lead times, but still higher than a pure climatology. The rainfall/runoff simulation still has memory of the first two days of QPF and the initial condition, which allows the simulation to be able to discriminate high flow events from non-events.

Issues, Recommendations and Actions from the Second RFC Verification Workshop
Salt Lake City, November 18-20, 2008

The NWS Hydrology Forecast Verification Team held a three-day verification workshop in November 2008 to present the progress made on verification activities at the RFCs, WFOs, and OHD, as well as in academia. All the workshop material is available online at http://www.nws.noaa.gov/oh/rfcdev/projects/rfcHVT_workshop2_agenda_presentations.html

Here are the key verification issues discussed during this workshop, including the team recommendations and proposed actions for future work.

1. <u>Users of forecast and verification products</u>

The team discussed the need to identify a few groups of customers of forecast products (including customers within NWS, such as hydrology managers) to define what kind of verification standard products would be meaningful for each group.
To do so, the team proposed that all the Service Coordination Hydrologists from the RFCs work with the RFC Verification Focal Point(s) to define these main groups of customers and propose a few verification metrics/graphics for each group. The verification team will have a meeting to discuss the proposed user groups and standards. Recommendations on verification standards for specific user groups will be given in the final verification team report.

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.

Service Coordination Hydrologists with RFC Verification Focal Points
NWS Hydrology Forecast Verification Team

2. <u>Verification strategies to effectively communicate verification information</u>

The team agreed that the NWS should provide at least three levels of sophistication in the verification information, from detailed statistics useful to forecasters, modelers and sophisticated users (e.g., with decision support tool that could directly ingest some verification results), to summary scores (e.g., green/red lights for various quality aspects) for the general public, along with explanations of such verification products.

Also the NWS should convey the forecast uncertainty information to customers, even with deterministic forecasts. One could have graphics with deterministic forecasts issued at different times to show how the forecast values are changing with lead time and get the user understand how difficult the situation is to predict. Also it would be useful to plot the probability forecast for a particular threshold (e.g. action stage) alongside the single-valued forecast; this plot will have a single line giving the probability of exceeding the selected threshold. In case a bias correction technique is available, a second line could be plotted for the bias-corrected probability. This approach will help users become familiar with probabilistic forecasts.

The team recognized the need to use normalized metrics to have meaningful inter-comparison results among basins and RFCs. For example, instead of using an absolute flow threshold value, the threshold value could be defined as one of the percentiles in the climatological record.

NWS Hydrology Forecast Verification Team

3. Raw model applications

Raw model forecasts, which are generated by the forecasting system with no/minimal human interactions at the time of the forecast issuance, are one of the meaningful references for verifying the operational forecasts for some specific users. These users include the forecasters, who want to evaluate how much value they add to the forecasts in various forecast situations, and the Hydrology Program managers, who evaluate what should be done to improve the forecasts. Depending on the users, the definition of raw model forecasts may vary. For inter-comparison purpose, it seems necessary to develop a unique national definition of raw model forecasts; each RFC could also define other raw model flavors to meet their own needs.

It was proposed that, at each RFC, the RFC Verification Focal Point(s) work with the Service Coordination Hydrologist to define how the raw model forecasts will be used and send their use cases and raw model definitions to the verification team. The verification team will have a meeting to discuss the different raw model definitions and use cases and will make a recommendation for the HICs.

RFC Verification Focal Points with Service Coordination Hydrologists
NWS Hydrology Forecast Verification Team

4. Future verification case studies for the RFCs

In several RFC presentations, forecasters mentioned their office choice concerning how many lead times of QPF is used to generate hydrologic forecasts. In some offices, a policy is in place to restrict the QPF lead times to very short lead time (e.g., 12 hours). This decision should be based on verification results to help evaluate in which situations more QPF values could be used to improve flow/stage forecasts. Therefore the verification team proposed to define a new verification study to be done at each RFC to inter-compare the quality of flow/stage forecasts based on different QPF lead times (0 QPF, 6 hours of QPF, 12 hours of QPF, 18 hours of QPF, … 10 days of QPF). Basically, this study requires producing in parallel and archiving various forecast runs, each one using a different QPF lead time, and evaluating in IVP the different sets of forecasts. For each forecast set, the verification metrics should be computed for each individual lead time for the whole time period, as well as sub-periods relative to specific weather or hydrologic conditions. Even if users may have different needs for these forecasts and thus accept to use forecasts with various quality levels, each RFC should be able to determine the optimal QPF lead times to meet the needs of their main users using such systematic verification case study.

Also the RFCs recognize the need to work on verification case studies using EVS to become more familiar with probabilistic verification. The EVS exercise that the team worked on in June and July 08 was a good introduction to EVS but each RFC needs to work with its own data. For example, EVS could be used to evaluate the long-term ESP forecasts (generated from climatological forcing inputs). Regarding the support on EVS, the RFCs should send all their questions to the verif-hydro list server to get some help from all the EVS moderators and users. Andrew Philpott at MARFC is now one of the moderators for EVS support.

RFC Verification Focal Points
NWS Hydrology Forecast Verification Team

5. Decomposition of flow error into timing error, peak value error, and hydrograph shape error

As one of the activities in the FY09 verification work plan presented during the workshop, HEP will work on decomposing the flow error into timing error, peak value error and hydrograph shape error. 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). Once the event pairing is done, standard statistics can be used (or if necessary, new metrics can be developed) to characterize the timing, peak value, and hydrograph shape error. Several techniques will be investigated, such as curve registration and object-based methods. The participants suggested starting with a simple manual pairing process and eventually including both an automated process and a manual process for the event pairing. The manual pairing process (similar to the STAT-Q tool used in calibration) should lead to an easier implementation of such functionality. This functionality will also be very beneficial to evaluating the quality of tide forecasts, for which the error decomposition would be very meaningful to users.

HEP Group at OHD

6. Verification training and communication

The team proposed that training on IVP and EVS, such as the software demonstration offered to the verification team, should be recorded as webinar to be offered at any time. GoToMeetings should also be held regularly to answer specific questions on software or interpretation of verification graphics.

Also the verification team will help COMET develop new verification modules. One module will present one verification case study with IVP and another one with EVS. These case studies will be defined once the verification team interim report has been developed.

The workshop participants recommended conducting another verification workshop in two years or so, 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.

The verif-hydro list server is a good tool to support the RFCs by sharing questions and solutions. A few rules need to be accepted to get efficient support:

- whenever someone wants to submit a question to the verification team, he/she should submit it to the list server, with in the title one of the four following categories: IVP and database issues; EVS; verification science; workshops and training;
- if you are one of the list server moderators, see if you can answer the question, or see if the question gets answered; if not, add the question to the list of unanswered questions, which will be revisited by the verification team at a future meeting.

HEP Group and HSEB at OHD
RFC Verification Focal Points
NWS Hydrology Forecast Verification Team

7. Define requirements for FY10 verification branch work

The current verification statistics compiled by the NWS Performance Branch 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.

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. These requirements should be developed by the verification team based on results from the case studies completed and ongoing. These requirements should be presented to the RFC and OHD management. The NWS Performance Branch should be engaged late in FY09 so that these requirements can become part of their FY10 workload.

NWS Hydrology Forecast Verification Team
NWS Performance Branch
RFC and OHD management

Appendix: Archiving Survey Results from the 13 RFCs
Results sent by 02/06/2008 to Julie Demargne

XEFS Edition – The entries highlighted in the table below identify the archive data type (in yellow) and its spatio-temporal scale (in green) necessary for calibration and parameter estimation of the XEFS components (for questions and further information, please contact Yuqiong Liu at yuqiong.liu@noaa.gov)

1. Description of available archived data

Variable, Data Type, Scale		RFC Name	Archiving frequency	Period of record		Format / Database	Time step	Fcst. / Re-fcst. Characteristics		Notes
				Start date	End date			Lead time for single-valued	Lead time for proba-bilistic	
P	Obs.	AB	Daily	6/24/94	Present	Xmrg/netcdf	1,6, 24 hr			QPE
R		AP - point	As received	2003	Present	Archive	15 min->1 day			Send to Archive
E	(Mean	AP – areal	Daily/as created	2006	Present	Archive	6 hr			
C	Areal,	CB	n times/day	2000	Present	Fastetc	6 hr			MAP - Can be moved
I	6-hourly	CB	n times/day	9/2005	Present	Xmrg files	1 hr			to Archive
P	or less)	CN	weekly	8/2007	Present	Archive db	1 hr			
I		CN	n times/day	2000	Present	Fastetc	6 hr			
T		LM	6 hr	1997	Present	DX	6 hr			
A		MA	daily	1996	Present	Text file	6 hr			MAP
T		MB	as required	1948/78	09/2004	Rax db/card	6 hr			MAP
I		NC	?	1948	2004	Card	6 hr			Upd'd annually
O		NE	daily	1948	Present	Archive/Card	6 hr			
N		NW – areal	daily	9/2005?	Present	Rax, files	6 hr			Dumpshef punchf.
		NW - point	obs/est value	?			1 hr, 6 hr, 24 hr			
		OH	hourly/contin's	1/1996	Present	Xmrg, Rax	1 hr, 1 day			
		OH - MAP		1948	2001	Card	6 hr			MAP
		SE - MAP	daily	5/1997	Present	Text, Rax/PC	6 hr			MAP
		SE - MAPX	daily	2/2005	Present	Text, Rax	6 hr			MAPX
		SE - Point	daily		Present	Dumpshef, DX	24 hr			
		SE - Point	As received		Present	Rax db	15 min-24 hr			
		WG	daily	01/1996	Present	xmrg	1 hr			
P	Fcst.	AB	Daily	10/3/94	Present	Netcdf	6 and 24 hr			QPF

PRECIPITATION		AB		4/3/96	Present	FMAP.d	6 hr	48 hrs		QPF
	(Mean Areal, 6-hourly or less)	AP – point	Daily/as created	2003	Present	MySQL	6 hr	72 hr		
		AP – areal	Daily/as created	2003	Present	MySQL	6 hr	72 hr		
		CB	n times/day	4/2002	Present		6 hr	3-5 days		
		CN	monthly	11/1996	Present	Text files	6 hr	?		FMAP
		LM	6 hr	1997	Present	DX	6 hr	72 hr		
		MA	3 times/day	8/2000	Present	Flat file	6 hr	48 hr		
		MA	daily	8/2001	10/2006	Card files	6 hr		7 days	EPP2
		MB	daily	7/2003	Present	Rax db/file	6 hr	1-3 days		IFP/OFS
		NC	Na							
		NE	daily	2007	Present	Archive	6 hr	1-2 days		Can repost n years
		NW – areal	daily	9/2005	Present	Rax, files	6 hr	1-10 days		Dumpshef punchf.
		NW – point	daily + updates	9/2005	Present	Flatfiles, Rax	6 hr	1-10 days		
		OH	daily	~2000	Present	Text files	6 hr	1-2 days		Extremes: longer ld t.
		SE	daily	6/2000	Present	Text, Rax/PC	6 hr	1 day		
		SE	daily	2/2005	Present	FMAP.d, Rax	6 hr	1 day		
		WG	daily	2003	Present	Text/QPS	6 hr and 24hr			QPF
TEMPERATURE	**Obs.**	AB	Daily		Present	Postgres	1 hr (mostly)			
		AP – areal	Daily/as created	2006	Present	Archive	6 hr			
	(Mean Areal, 6-hourly or less)	AP - point	As received	2003	Present	Archive	15 min -> 1 day			
		CB	n times/day	2/2000	Present	Fastetc	6 hr			MAT
		CN	n times/day	10/2007	Present	Fastetc	6 hr			MAT
		LM	na							
		MA	Daily - seasonal	1/2001	Present	Text file	6 hr			Not in summer
		MB	As required	1948/78	Present	Card	6 hr			MAT
		MB-pt Mx Mn	daily	11/1994	Present	Rax db, text f.	24 hr			IFP/OFS
		NC	?	1948	2004	Card	6 hr			Upd'd annually
		NE		1948	Present	Rax db/Card	6 hr			
		NW - areal	Daily	9/2005	Present	Rax, files	6 hr			Dumpshef punchf.
		NW-pt Mx Mn	?	?						
		OH	hourly	3/2005	Present	Archive	1 hr			
		OH - MAT		1948	2001	Card	6 hr			
		SE	na							
		WG	na							
TEMPERA	**Fcst.**	AB	na							Use model fcsts
		AP	na							
	(Mean Areal, 6-hourly or less)	CB	n times/day	6/2003	Present	Fastetc	6 hr	5-10 days		
		CN	monthly	11/1996	Present	Text files	6 hr			FMAT
		LM	na							
		MA	Upon issuance	10/2000	Present	Text files	6hr	5 days		Station only
		MA	Daily	8/2001	10/2006	Flat file	6hr		7days	EPP2

T		MB-pt Mx Mn	daily	7/2003	Present	Rax db, text f.	24 hr	6 days		IFP/OFS
U		NC	na							
R		NE	daily	2007	Present	Archive db	6 hr			Can repost n years
E		NW – areal	daily	9/2005	Present	Rax, files	6 hr	1-10 days		Dumpshef punchf.
		NW-pt Mx Mn	daily	9/2005	Present	Flatfiles, Rax	24 hr	1-10 days		
		OH	na							
		SE	na							
		WG	na							
F	**Obs.**	AB	na							
L		AP	Conv'd. from stg	2003	Present	Archive	15 min -> 1 hr			
O	**(6-hourly**	CB	n times/day	Start real t.	Present	Fastetc	1 hr			
W	**or less &**	CN	monthly	11/1996	Present	Text file	1 hr			
	daily)	CN	weekly	8/2007	Present	Archive db	1 hr			
		LM	Na							
		MA	Daily	1/1999	Present	Text file	15 min			USGS data
		MB	As required	10/1948	Present	Card	1 day			calibration
		NC	?	1948	2004	Card	1 day			
		NE	?	1948	~2000	Card	6 hr			
		NW	?							
		OH	na							
		SE	daily	3/2005	Present	Text, Rax db	6 hr			
		WG	daily	11/2007	Present	Rax	6 hr			
F	**Fcst.**	AB	na							
L		AP	na							
O	**(6-hourly**	CB	n times/day	1/2000	Present	Fastetc	1 hr	2-5 days		Routine model out.
W	**or less)**	CN	1-4 times/day	1/2005	Present	Text files	6 hr	5 days		RNORR1RSA
		CN – volume	Monthly	1/1997	Present	Fastetc	Volume		120 days	Water supply
		LM	Na							
		MA	Daily	?			6 hr	5 days		
		MA	Daily	8/2001	Present	CS/HS files	6 hr		7 days	Based on EPP2
		MA	Weekly	2002	Present	CS/HS files	6 hr		30 days	AHPS products
		MB – vol.	Monthly	10/1948	Present	Fastetc, db				Water supply
		MB	Monthly	10/2005	Present	Files	Volume		90d-6mo.	ESP
		MB (SHEF)	Daily (yr/season.)	2003	Present	Db, text files	6 hr	5d.-26d.		IFP/OFS
		NC	Monthly	2002	2007	CS files	6 hr		90 days	ESP
		NE	Na							
		NW – regress	3/mo. Water sup	1995	Present	Westwide doc	various			
		NW – ESP	Weekly	2004	Present	Flatfiles	various			
		OH	daily	1/2000	Present	Text file	6 hr	6d. - 28d.		
		SE	daily	2/2005	Present	Text, Rax db	6 hr	5 days		

		WG	daily	11/2007	Present	Rax	6 hr			
F L O W	Re-fcsts	CB	As needed			Fastetc	1 hr			
S	**Obs.**	AB	Daily	10/1993	Present	Postgres	15 min -> 1 hr			1993-1999 6 hr
T		AP	As received	2003	Present	Archive	15 min -> 1 hr			
A	**(6-hourly**	CB	?							
G	**or less)**	CN	monthly	1/2001	Present	Text f., fastetc	15 min			
E		CN	weekly	8/2007	Present	Archive db	1 hr			
		LM	Hourly	1997	Present	DX	1 hr			
		MA	Daily	8/1996	Present	Text file	15 min			
		MB	Daily	11/1994	Present	Rax db	varies			
		NC	?							
		NE	Daily	1998	Present	Rax db	1 hr			Data in other formats
		NW	?							
		OH	Continuous	1/2004	Present	Rax	1 hr			
		SE	As received	1/2001	Present	Rax	15 min-24 hr			
		WG	Continuous	2/2000	Present	Rax	1 hr			
S	**Fcst.**	AB	Daily		Present	Postgres	6 hr			
T		AB	daily	8/1995	Present	GXSETS ascii	6 hr			
A	**(6-hourly**	AP – FE	Daily/as created	2003-05	Present	Archive	6 hr or 1 day	1-3 days		
G	**or less)**	AP - FZ	Daily/as created	2003-05	Present	Archive	6 hr or 1 day	1-4days		
E		CB	During events	2003	Present	Files	6 hr, 1 hr			Official fcsts
		CN	1-4 times/day	12/2004	Present	Text file	6 hr	5 days		RNORR1RSA
		LM	6 hr	5/2003	Present	Ax	6 hr			
		MA	Daily	11/2000	Present	Text file	6 hr	5 days		w/ OFS snapshot
		MB	Daily	1/1983	Present	Rax db, text f.	6 hr	5d.-26d.		IFP/OFS
		MB	Monthly	10/2002	Present	Text f., ESG f.	Peak value		90d-9mo.	ESP
		NC	Monthly	2002	2007	CS files	6 hr		90 days	
		NC	Daily	2005	2007	SHEF	6 hr	7-14 days		
		NE	daily	1998	Present	Rax db	6 hr	2 days		
		NW	daily +	7/2003	Present	Rax, flatfiles	6 hr (some 1 hr)	1-10 days		
		OH	continuous	1/2004	Present	Rax	6 hr	5-7 days		
		OH	daily	1997	Present	Text file	6 hr	3-5-7 days		
		SE	Daily	2/2005	Present	Text, Rax db	6 hr	5 days		
		WG	As issued	11/2003	Present	Rax	6 hr	3-5 days		
	Reservoir Obs.	MB - reservoir	daily	11/1994	Present	Rax db	24 hr			
	Reservoir	CN –	1-4 times/day	12/2004	Present	Text file	6 hr	4 days		RNORR1RSA

	Fcst.	reservoir								
		elevation								
		CN – full	1-4 times/day	12/2004	Present	Text file	6 hr	5 days		RNORR1RSA
		natural flow								
		MA –	Daily/when iss'd	2007	Present	Text file	6 hr	2 days		Pool elevations
		reservoir								
		MB -	When issued	2004	Present		6 hr	5 days		IFP/OFS
		reservoir								
	MPE	CB	As created	11/2003	Present	xmrg	1 hr			
	MODs	AB	?	12/8/03	Present	ASCII				
		AP	Daily/as created	2007	Present	Flat files	6 hr			
		CB	n times/day	1/2000	Present	Fastetc				
		CN	6 months	1999	Present	Text files				
		LM	6 hr			DX				
		MA	Continuous	1/2005	Present	Text file	6 hr			
		MB	Daily, weekly	10/2005	Present	Text f./fs5files				
		NC	Daily	2005	2007	Ascii				
		NE	na							
		NW	Daily +	10/2003	Present					
		OH	Daily	5/2005	Present	Text file				
		SE	Daily	1/2006	Present	Text, RAX				
		WG	daily	1/2006	Present	Gzip/tar	6 hr			
	Rating	AB	na							
	Curves	AP	Manually entered	2003	Present	Archive				
		CB	As updated	1/2000	Present	Fastetc				
	As up-	CN	As updated	5/2000	Present	Fastetc				
	dated	LM	As updated	~1995						
		MA		5/2003	Present					w/ OFS snapshot
		MB	As updated							
		NC	na	8/2006	Present					
		NE	na							
		NW	As updated	8/2006	Present					
		OH	Intermittently	1998	Present	Text file				
		SE	na							
		WG	na							
	OFS Files	AB fs5files	daily							On pc170 machine
		CB fs5files	3 times/week	4/2004	Present	Files				
		CN	3 times/week	2004	Present	Files				
		MA	Weekly	1/2005	Present					OFS snapshot
		NC	daily	2005	2007	Binary	6 hr			

		NW	daily +	10/2003	Present					
		SE fs5files	monthly	1/2006	Present	Binary, Rax	6 hr			
		WG fs5files	4 times/day	12/2005	Present	Gzip/tar	6 hr			
	Verifica-tion Statistics	AB – stage	?	1/1997	Present					
		AP – FE	Monthly	2005	Present	Archive				From May to Oct.
		AP – FZ	monthly	2005	Present	Archive				From May to Oct.
		CB	na							
		CN – RVF	monthly	5/2001	Present	Text files, Excel	6 hr	72 hr		
		CN – study	once	12/2005	01/06	Archive	6 hr	72 hr		WR case study
		LM	monthly	7/2005	Present	Ax	6 hr			
		MA	monthly	2001-05	Present	Archive db	6 hr			
		MB - RVF	monthly	1/2005	Present	Text files	6 hr	3 days		
		NC	na							
		NE	na							
		NW	na							
		OH	na							
		SE	monthly	4/2001	Present	Text, DX/Rax	6 hr	3 days		
		SE	monthly	2/2005	Present	Rax db	6 hr			
		WG	na							

2. Description of current archiving processes and issues

Consider time/speed issues, storage issues, memory issues, process flexibility/complexity, and comment as necessary.

2.1. Data management

- **Archive database**

APRFC: - QC issue
- we need a way to have offsets to transmitted data be applied before sending to the Archive Database.

CBRFC: We store all of our data in our own database (fastetc), which the archive db was somewhat based on. Data is moved to the archive database as needed for verification (national verification data is moved each month).

CNRFC: We store most of our data in fastetc. Data feed (observed stage and precipitation) to the archive db started August 2007.

LMRFC: Our Archive database is unstable and goes down frequently (every few months). Hopefully the new install will be more robust.

MARFC:

- All products received on AWIPS and handed to the operational shefdecode are passed to the raw decoder
- Once per day, processed OFS values (MAP, MAT, FMAP, SWE, MAPX) are run through the processed decoder
- Concern is the length of time DB restores will take with increasing amounts of data stored in RDBMS
    - These occur with builds when DB structure changes occur and when the system crashes

MBRFC: There are known performance issues (time/speed, memory issues) with the Archive DB on the RAX. Not enough disk space for all the flat files that should be in the flat file archive. Have many files stored in locations other than the RAX that were never moved to the RAX due to lack of space and inadequate backups and recovery for RAX (its single point of failure).

NCRFC: Dual archive databases run. Verification is being run on a surplus machine (SAD) with better performance. We have recently split our data stream to provide individual data feeds to the Postgres and SAD archive databases. AX database is currently near disc capacity.

NWRFC:

Archive Database Deficiencies

- Available storage is inadequate
- Posting data via the Raw Shef Decoder is too slow (cannot keep up with incoming data)
- Posting speed appears to be HEAVILY dependent on file que (is decoder reading file list into memory?)
- Processed Shef decoder cannot process data unless all 7 shef parameters are specified (i.e it won't assume last two characters (ZZ) like the raw decoder)
- Does not store sub-hourly time-step data in correct table
- Archive database filled up

RAX interactions with software (VFYRUNINFO, IVP, IVPBATCH, DATVIEW)

- vfyruninfo can take 10-15 minutes for the gui to come up
- Memory issues occur when multiple plots are created using IVP (even with -m option)
- Generating a time-series plot from scatter plot display takes an inordinate amount of time (several minutes)
- Running National Stats package takes significantly longer (hours) to complete in version 8.2
- Datview will experience memory fault when viewing as few as two years of data

SERFC: Data retrieval is sometimes slow. Disk space needs to be increased.

WGRFC: small and slow (Note: with the recent upgrade in Summer 08, the slowness is remedied). It often takes 28 hours to pair 2 days worth of VAR, SSHP, and DHMS hourly forecasts. The lack of size also ‘’scares” us on fully archiving precipitation data. We will do it, but that was one of the main reasons for not doing so in the past.

- **Files:**
    - **Format**
    - **Directory structure**
    - **Naming convention**
    - **Metadata**

ABRFC:

- Format - Need to define the format that all offices will save products in. Will it be xmrg for gridded data?
- Directory structure - If we use the ax, this should be pretty much decided.

CBRFC:

- We are not using the archive machine for our archive files at this time.
- We archive qc'd observed hourly flows.
- We store model simulated and adjusted flows (both past and future) to our database after each run
- We have shef files of our official stage forecast products issued during events that can easily be stored to the database.
- We store all of the old ratings and shifts in our database so we can convert between stage and flow as needed.
- We have the ability to make re-forecasts. These are done as needed for specific purposes.

MARFC:

- Format – mostly tarballs
- Directory structure – Latest RAX-provided structure
- Naming convention – Follows naming convention used by OHD when they had archive responsibilities
- Metadata – Added a "Notes.txt" file at the "YEAR" level and in the directory below where tweaks/changes/corrections needed
- Methods – Using the "archive_move.pl" and "archive_monitor.pl" approaches distributed by J. Meyer

MBRFC:

- Format: currently no issue with this for what is electronically stored. Currently only current rating is available. History of rating shifts is stored electronically for each OFS site but history of base ratings is available only in hardcopy in station files.
- Directory structure: currently no issue with this
- Naming convention: currently no issue with this
- Metadata: with build ob8.3 will finally be able to easily sync meta-data between IHFS DB and Archive DB.

NCRFC:

- Directory structure: RVF files are stored on the AX under the flatfiles directory.
- Naming convention: ESP files are named by id.id.datatype.interval.type (e.g. WAPI4.WAPI4.SQME.24.OBS)

NWRFC:

(1) Directory structure – Flat file structure is too rigid!

SERFC:

- Metadata: Looking forward to AWIPS Build OB8.3 to easily sync meta-data between IHFS DB and Archive DB

2.2. Data Quality Control

ABRFC: Done on a daily basis, and results sent to ax. Also done once again when monthly verification stats are generated.

APRFC:

This is to me the biggest issue. I don't know how other RFC's are dealing with this, but we send all incoming data to the Archive Database as it is ingested. The data is QC'd by varying methods, including graphically through Time Series display, and the QC'd data does NOT get written to the Archive as either raw or processed data. Several points of view exist as to how QC'd data get into the archive. I don't care.

My problem is that, as a result, I spend several DAYS at the beginning of each month RE-QC'ing the data for verification. Also, since data that are set to missing in Time Series DO NOT set a missing value in the Time Series Postgress database, I have no way of easily identifying some of the bad data that reaches the Archive.

There needs to be a simple GUI (like the Time Series graphical display) to allow simple editing of Archive data. The current process of viewing and editing the Archive is not good.

CBRFC:

Quality control of model inputs (precipitation, temperature and stage/flow data) is done multiple times per day during routine operations. First pass is with automated programs and second pass is manual. This qc'd data is what is archived in our database and fed to the model.

CNRFC: Model inputs (precipitation, temperature and stage/flow data) are Qc'd before each model run. Precipitation and temperature Qc is done with Mountain Mapper graphical interface. Stage/flow Qc done with IHFS Hydro Time Series display. Qc done again when monthly RVF verification statistics are generated.

LMRFC: Bad data is obvious sometimes; time is taken to remove it.

MARFC: QC stage data performed by operational staff via Time-Series app; SHEF messages with missings sent to raw decoder where stages were deleted. Verification QC time greatly reduced as a result of these actions.

MBRFC: Data in archive db's raw shef values tables needs to be qc'd as part of the pairing process. Primarily stage data has problems but some data already in the vfypairs table has been qc'd and if pairing is not rerun for a station/dates already in the vfypairs table, should be o.k.

NCRFC: Stages, precip and temps QC'd daily by forecasters, corrections sent to AX shefdecoder. ESP output QC'd to extent possible for model problems, bad data, etc.

NWRFC: we need better tools to edit and display data.

SERFC: DatView provides a decent way to correct Raw Observed Data. Process can be time consuming.

WGRFC: our DOH developed a tcl script that allows for easy (real easy) QC of data using the Datview program. This is no issue.

2.3. Back up

ABRFC: None currently, DEFINITELY need this!

CNRFC: run_PGbkups is done every week. Dump file: adb_ob72rsa.dump. Was 22941586 bytes in size on 08/28/2007, and 39566479 bytes on 12/12/2007, and 47631962 bytes on 01/23/2008. Archive database is growing due to initiation of data feed. Dump files stored on remote hard drive.

MARFC:

- Postgres tables backed up daily and copied to non-RAX file system
- Weekly multi-file tape
- RAX recently died and these two backup methods provided a nearly full restore of files and directories
    - Gaps filled using short-term archive of information on AWIPS (14-30 days) to fill in from point of tape and DB backup forward

MBRFC: Ensuring local back-up procedures are in place and working. RAX hardware is single point of failure. Archive database and flat files that are on the RAX are at risk.

SERFC: The RAX hardware is a single point of failure. SERFC's RAX system has failed twice completely since it was installed in August 2003. Complete backup and recovery procedures need to be established for the entire system (not just data). More than 2 spare systems were needed at NLRC. At SERFC cron_PGbkups is run daily and the resultant dump file is backed up on the RAX and an external system. Flatfiles are backed up monthly.

WGRFC: flat files are backed up to DVD 1 per year. The database is archived to a flat file using Julie Meyer's scripts weekly.

2.4. Foreseen problems when interacting with IVP (note: issues are contingent on very limited experience with the IVP ob8.2 presented at the RFC verification workshop in August 2007)

ABRFC: None, just need to get more comfortable with it.

APRFC: at this point, very early for us, data QC and software documentation

CBRFC: the only issue we have is making sure that we move everything we need over to the archive machine. Since we have not fully populated the archive database or it's flat file archive directories, we have not experienced any slowdowns when running IVP. However, we do have concerns that if we did we might run into problems similar to NWRFC.

CNRFC: Takes about 45 minutes to run monthly RVF statistics that's sent to OCWWS. No concerns with speed running IVP yet.

MBRFC: Current RAX has slowness issues, remembering to have patience is key. Slowness issue varies from RFC to RFC. Also there is a problem with the RAX DB at two RFCs… NCF has not been able to pinpoint the source of the problem (hardware vs. software).

NCRFC: Running on large samples is extremely slow. We currently run pairing and stats in batch mode for one month at a time. Graphics are generated using excel due to time constraints on AX. We can use IVP to generate graphs for a single location for a one month period, but any samples larger than this start to slow down performance significantly.

NWRFC: It is pretty obvious that performance suffers dramatically with large datasets.

SERFC: Lack of an understanding of the display may be an issue.

WGRFC: meaning and understanding of verification statistics.

2.5. Foreseen problems when interacting with EVS (note: issues are contingent on very limited experience with the EVS beta 1 prototype presented at the RFC verification workshop in August 2007)

ABRFC: None, just need to get more comfortable with it.

APRFC: at this point, ground zero, data QC and software documentation

SERFC: Lack of an understanding of the display may be an issue.

WGRFC: learning new statistics terminology for probabilistic forecasts and verification