Real-time Verification of Short-term Probabilistic Streamflow Forecasts

Is it possible?

Bisher Imam
University of California, Irvine

Holly Hartmann
University of Arizona
**Limitation and Research Needs for ESP**

“*The AHPS approach to quantifying uncertainties in operational forecasts must be articulated*”

**Limitations and Research Needs for Verification in the NWSRFS**

“*Unlike meteorological forecasts, little is known about hydrologic forecasts and actual river forecast skills*”
Objectives

• Identify suitable measures for real-time ESP forecast verification
• Propose operational procedure for ESP forecast verification
• Propose examples of possible screens that can be integrated into the system
ELEMENTS OF FORECAST VERIFICATION

- Event
- Verification Measures
- Attributes
- Forecast Type
Focus on Forecasters
Hindcasts: 1953-1993 → 45 Years
Ensembles: 1952-1953 → 46 Ensembles
Procedure

NWSRFS System

Observed

/calb/historical files (MAP,MAT,QME,QINE)

Historical ascii files

Climatologic probabilities

R Statistical Package

R verification package

Ensemble Simulations

R Tesgen

RTI-tstool

NWSRFS's Ensemble files in (/ens/output/verification)

Ensemble ascii files

Forecast Probabilities (hindcast mode)

Center for Hydrometeorology and Remote Sensing, University of California, Irvine
ELEMENTS OF FORECAST VERIFICATION

- Event
- Forecast Type

Attributes

Verification Measures
- Box Plot
- Spread
- Histogram
- Q-Qplots
- Scatter Plots
- Association
- Scatter Plots with Histograms
- Combined
- Bivariate Histograms
- Exploratory
- Bias
- Mean Square Error
- Accuracy
- Uncertainty
- Summary Statistics
- Correlation Coefficient
- Association
- Calibration Refinement: Observation | Forecast
- Joint Distribution
- Marginal/Conditional
- Distribution Based
- Murphy and Winkler
- Likelihood-base rate: Forecast | Observation
Exploratory Approaches; The Ensemble Plots

Frcst-Yr 1953

Frcst-Yr 1965

Center for Hydrometeorology and Remote Sensing, University of California, Irvine
Multiple Forecasts
Consistency: Error and Spread

\[ |\mu_{\text{Ens}} - \text{Obs}| \]

\[ \sigma_{\text{Ens}} \]
Normalized Error/Spread Plot
Larger Data Set: All Ensembles Considered
What is a reasonable forecast variable?
Distribution oriented measures

CR Factorization

\[ h(f, o) = y(o \mid f)p(f) \]

LBR Factorization

\[ h(f, o) = r(f \mid o)t(o) \]

\[ f_i(q^*) = P\{Q_i \leq q^* \mid \alpha_i\} \]

where \( f_i \) is the probability forecast, and \( \alpha_i \) is the initial condition. Observation variable \( o_i(q^*) \) as

\[ o_i(q^*) = \begin{cases} 
1 & \text{if } Q_i \leq q^* \\
0 & \text{if } Q_i > q^* 
\end{cases} \]
RELIABILITY:
Forecast probabilities for given event match observed frequencies of that event (with given prob. forecast)

RESOLUTION:
Occurrence and non-occurrence of event is well resolved by forecast system
Random Forecast

Overall Performance Measures
- Brier Score (BS) = 0.3306
- Brier Score - Baseline = 0.2497
- Skill Score = -0.324
- Reliability = 0.08339
- Resolution = 0.0025
- Uncertainty = 0.2497

Reliability

BS-Decomposition
Real Forecast (Augmented 15 Days)

Overall Performance Measures
- Brier Score (BS) = 0.1431
- Brier Score - Baseline = 0.25
- Skill Score = 0.4276
- Reliability = 0.007992
- Resolution = 0.1149
- Uncertainty = 0.25

Center for Hydrometeorology and Remote Sensing, University of California, Irvine
Thresholds and Probability Range Selection

threshold = 0.1 Flow = 78

threshold = 0.2 Flow = 177

threshold = 0.3 Flow = 367

threshold = 0.4 Flow = 600

threshold = 0.5 Flow = 839

threshold = 0.6 Flow = 1046

threshold = 0.7 Flow = 1351

threshold = 0.8 Flow = 2060

threshold = 0.9 Flow = 3128
All thresholds together (Also applicable to individual thresholds)

Overall Performance: Historical Simulation
- Brier Score (BS) = 0.05767
- Brier Score - Baseline = 0.25
- Skill Score = 0.7693
- Reliability = 0.002273
- Resolution = 0.1946
- Uncertainty = 0.25

Overall Performance: Observation
- Brier Score (BS) = 0.1432
- Brier Score - Baseline = 0.25
- Skill Score = 0.4273
- Reliability = 0.007799
- Resolution = 0.1146
- Uncertainty = 0.25

BS-Decomposition

Attribute Diagram

BS-Decomposition
**ROC Curve**

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p</td>
<td>n</td>
</tr>
<tr>
<td>Y</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>N</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
<tr>
<td>Total</td>
<td>P</td>
<td>N</td>
</tr>
</tbody>
</table>

False Positive rate = FP/N = False Alarm Rate
True Positive Rate = TP/P = Hit Rate
Precision = TP/(TP+FP) = Positive Predictive value
Accuracy = (TP+TN)/(P+N)
Specificity = TN/(FP+TN)

Threshold = 0
Always make a positive forecast

Threshold = 1
Never make a positive forecast
Combining Information

**ROC Plot Historical Simulation**

- Non-exceedence Threshold: 0.7
- Overall Performance: Observation
  - Brier Score (BS) = 0.1657
  - Skill Score = 0.2113
  - Reliability = 0.0187
  - Resolution = 0.0631
  - Uncertainty = 0.2102
- Overall Performance: Historical Simulation
  - Brier Score (BS) = 0.0698
  - Skill Score = 0.696
  - Reliability = 0.001809
  - Resolution = 0.1481
  - Uncertainty = 0.2102

**BS-Decomposition**

- Brier Score
- Reliability
- Resolution
- Skill Score

**Attribute Diagram**

- Observed Relative Frequency, \( T \)
- Historical Simulation
- No resolution

**BS-Decomposition**

- Brier Score
- Reliability
- Resolution
- Skill Score
Screen Shot (15 days)

Center for Hydrometeorology and Remote Sensing, University of California, Irvine

**ROC Plot Historical Simulation**
- Non-exceedence Threshold: 0.7
- Overall Performance: Observation
  - Brier Score (BS) = 0.1657
  - Brier Score - Baseline = 0.2102
  - Skill Score = 0.2113
  - Reliability = 0.0187
  - Resolution = 0.0181
  - Uncertainty = 0.2102
- Overall Performance: Historical Simulation
  - Brier Score (BS) = 0.06388
  - Brier Score - Baseline = 0.2102
  - Skill Score = 0.696
  - Reliability = 0.001809
  - Resolution = 0.1481
  - Uncertainty = 0.2102

**BS-Decomposition**
- Brier Score
- Reliability
- Resolution
- Skill Score

**Attribute Diagram**

**BS-Decomposition**
- Brier Score
- Reliability
- Resolution
- Skill Score
For smaller Data Set

ROC Plot Historical Simulation

- Non-exceedance Threshold: 0.7
- Overall Performance: Observation
  - Brier Score (BS) = 0.1189
  - Brier Score - Baseline = 0.2106
  - Skill Score = 0.4353
  - Reliability = 0.02494
  - Resolution = 0.1166
  - Uncertainty = 0.2106
- Overall Performance: Historical Simulation
  - Brier Score (BS) = 0.04961
  - Brier Score - Baseline = 0.2106
  - Skill Score = 0.7645
  - Reliability = 0.01962
  - Resolution = 0.1806
  - Uncertainty = 0.2106

ROC Plot Observation

BS-Decomposition

Attribute Diagram

BS-Decomposition
<table>
<thead>
<tr>
<th>Measured Quality</th>
<th>Base dist.</th>
<th>Measure</th>
<th>Definition (Murphy 1996 and 1997)</th>
<th>Formulae (Based on Bradley (2003))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>$p(f)$</td>
<td>ME</td>
<td>Mean error. Difference between the mean forecast probability and the climatological probability of the specific threshold.</td>
<td></td>
</tr>
<tr>
<td>Unconditional</td>
<td>$p(o)$</td>
<td></td>
<td></td>
<td>$ME = \mu_f - \mu_o$</td>
</tr>
<tr>
<td>Accuracy</td>
<td>$h(f,o)$</td>
<td>MSE</td>
<td>Mean square error. Overall degree to which forecast corresponds with observations.</td>
<td>$MSE = (\sigma_f^2 + \mu_f^2) + \mu_o(1 - 2\mu_f)\mu_{f=1}$</td>
</tr>
<tr>
<td>Skill or Relative</td>
<td>$h(f,o)$</td>
<td>SSMSE</td>
<td>A normalized measure of accuracy. Also known as the Brier Skill Score.</td>
<td>$SS_{MSE} = 1 - (\frac{\sigma_f^2 + \mu_f^2}{\sigma_o^2} + \mu_o(1 - 2\mu_f)\mu_{f=1})$</td>
</tr>
<tr>
<td>Accuracy</td>
<td>$h(f,o)$</td>
<td>$\rho_{f,o}$</td>
<td>A measure of association between forecasts and observations. Potential skill of the perfectly reliable forecast (See definition of reliability)</td>
<td>$\rho_{f,o} = \frac{\text{cov}(f,o)}{\sigma_f \sigma_o} = \frac{\mu_o}{\sigma_f} \left( \frac{\mu_{f=1} - \mu_f}{\sigma_f} \right)$</td>
</tr>
<tr>
<td>Reliability/ Type 1</td>
<td>$y(o,f)$</td>
<td>REL</td>
<td>Degree of correspondence between the observations associated with a given forecast and the forecast. Also known as Type 1 conditional bias. In meteorology it is also known as Calibration.</td>
<td>$REL = E_f (\mu_{o</td>
</tr>
<tr>
<td>Conditional Bias</td>
<td>$p(f)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution</td>
<td>$\sigma_f$</td>
<td>RES</td>
<td>Degree of spread of observations around the conditional mean of observation for a given forecast.</td>
<td>$RES = E_f (\mu_{o</td>
</tr>
<tr>
<td>Sharpness</td>
<td>$p(f)$</td>
<td>$\sigma_f^2$</td>
<td>Degree to which probability forecasts approach 0 and 1.</td>
<td>$\sigma_f^2$</td>
</tr>
<tr>
<td>Concurrency</td>
<td>$y(o,f)$</td>
<td>$\sigma_f^2$</td>
<td>Concurrence between reliability and resolution</td>
<td>$MSE_{CR} = \sigma_f^2 - REL - RES$</td>
</tr>
<tr>
<td>DISCR</td>
<td>$p(f)$</td>
<td></td>
<td></td>
<td>$SS_{CR} = 1 - \frac{MSE_{CR}}{\sigma_f^2} = \frac{RES}{\sigma_o^2} - \frac{REL}{\sigma_o^2}$</td>
</tr>
<tr>
<td>Discrimination</td>
<td>$r(f,o)$</td>
<td>DIS</td>
<td>Degree of deviation between the conditional mean for forecast from the mean of forecast.</td>
<td>$DIS = E_o (\mu_{o</td>
</tr>
<tr>
<td>Type 2 conditional</td>
<td>$r(f,o)$</td>
<td>B2</td>
<td>Degree of correspondence between the mean of forecast conditioned by observation and the mean of observations.</td>
<td>$B2 = E_o (\mu_{f</td>
</tr>
<tr>
<td>bias</td>
<td>$r(f,o)$</td>
<td></td>
<td></td>
<td>$MSE_{LBR} = \sigma_f^2 + B2 - DIS$</td>
</tr>
<tr>
<td>Relative Type 2</td>
<td>$r(f,o)$</td>
<td></td>
<td></td>
<td>$SS_{LBR} = 1 - \frac{\sigma_f^2}{\sigma_o^2} + \frac{MSE_{LBR}}{\sigma_o^2} - \frac{DIS}{\sigma_o^2}$</td>
</tr>
</tbody>
</table>
Forecasting is a Demanding Task

I. Preparation
- Morning Time (6-8AM)
  - Use Data Ingestion Tool
    - Mostly Gauges
    - GOES, ALERT
  - QC Precipitation
    - Western RFCs do not use RADAR
  - Mountain Wrapper
    - WESTERN_REGION
    - APPLICATION
  - Raw Data Archived
    - Automated Acquisition
    - Flat Files
  - Stream flow data
  - HAZ Prepares QPFs
    - Early Morning
  - Rating Curves
    - downloaded daily
    - updated twice a week

II. Process data
- Simple and fast automated
- Create MAP/IMAT
- Process MAP and IMAT for entire week

III. Situation Assessment
- Forecast Groups
  - Forecasts a day
  - Reservoirs
  - 4 Forecasts a day
- Forecast Groups
  - Load selection
    - Begin
    - Blend Period
  - MDS
  - Rating Curves
    - Offsets
  - Reservoir Issues
- Info Forecast
- SHEF Encoded product addressed to specific users
- River Guidance
- Web Products

IV. IFP_MAP
- Forecasting (CNRFC)
  - QPF Adjustment
  - 24 hrs
  - 4 Forecasts a day

V. ESP
- Segment Command
  - 5 day volume for 20 days
  - Generated automatically using Scripts

Get data for past day or length

Get e-mail about changes in reservoir operations from RES Operators (15 min)

Obtain 7 Days Forecast Temperature
(Mostly MOSS) SHEF.b (Station list)
- Max
- Min
- Twice a day

Collected in Database
- QC by CH through automated process
- Runs hourly

Observed Max-Min generated from 6 Hourly Temperature

Use Data Ingestion Tool
- Mostly Gauges
- GOES, ALERT

Mark stations as questionable
- Force it to be used
- Estimate value for stations that are marked
- Decisions require knowledge of individual gauges

Create MAP/IMAT
- Process MAP and IMAT for entire week

Look at stations
Wells, 2005. The objective of administrative verification of deterministic river stage forecasts is to determine:

1. How does the performance of the actual forecasts compare to the performance of persistence forecasts?
2. How does the forecast performance change with lead time?
3. How does the forecast skill change with time?

For ESP, no archive of forecasts exist because of:

1. Recent implementation and continuing evolution of ESP procedures at RFCs
2. Lack of archival procedures of actual ESP forecasts in their numeric (ensemble) format.
While the Inclusion of a verification subcomponent in AHPS NWSRFS is recommended, there is a pressing need for a long-term strategy and maintenance of forecast archive for future verification and NWSRFS evaluation.