Hydrologic Forecast Verification

Prepared by Satish Regonda,
Julie Demargne and D.-J. Seo

Hydrologic Ensemble Prediction (HEP) Group
Office of Hydrologic Development
NOAA/National Weather Service

RFC Short-Term Ensemble Workshop, November 30, 2006
Sources of this Presentation

- Brown, B.G., (2001), Verification of Probability Forecasts at Points, *WMO QPF Verification Workshop, Prague, Czech Republic*
- Ebert, B., (2003), Verification of Nowcasts, *WWRP Nowcasting Training Workshop, 3-14 November 2003, INMET*
- Ebert, B., (2005), Verification of Ensembles, *TIGGE workshop, 1-3 March 2005, ECMWF*
- Hartmann, H.C., (2006), Hydrologic Forecast Verification
- Hagedorn, R., (2006), EPS Diagnostic Tools, *ECMWF Training Course, Reading*
- Seo, D.-J., (2005), Tutorial Examples of Reliability Diagram and Relative Operating Characteristic from Short-Term Ensemble Prediction, *DSST Ensemble Verification*
- Welles, E., (2004), Hydrology and Verification, *DOH/RDM Conference*

*All these presentations are available either online or on OHD server*
Types of Forecasts

Deterministic

Categorical

Probabilistic

“Today’s high will be 76 degrees, and it will be partly cloudy, with a 30% chance of rain.”

How would you evaluate each of these?

Source: Hartmann (2006)
So Many Verification Measures!

- Quantifiable measures

<table>
<thead>
<tr>
<th>Deterministic</th>
<th>Categorical</th>
<th>Probabilistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>Hit Rate</td>
<td>Brier Score</td>
</tr>
<tr>
<td>Correlation</td>
<td>Surprise rate</td>
<td>Ranked Probability Score</td>
</tr>
<tr>
<td>RMSE</td>
<td>Threat Score</td>
<td>Score</td>
</tr>
<tr>
<td>Standardized</td>
<td>Gerrity Score</td>
<td>Rank Histogram</td>
</tr>
<tr>
<td>RMSE</td>
<td>Success Ratio</td>
<td>Distributions-oriented Measures</td>
</tr>
<tr>
<td>Nash-Sutcliffe</td>
<td>Post-agreement</td>
<td>Resolution</td>
</tr>
<tr>
<td>Linear Error in Probability Space</td>
<td>Percent Correct</td>
<td>Reliability</td>
</tr>
<tr>
<td>Skill scores</td>
<td>Pierce Skill Score</td>
<td>Discrimination</td>
</tr>
<tr>
<td></td>
<td>Gilbert Skill Score</td>
<td>Sharpness</td>
</tr>
<tr>
<td></td>
<td>Heidke Skill Score</td>
<td>Relative value</td>
</tr>
<tr>
<td></td>
<td>Critical Success index</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hannsen and Kuipers Score</td>
<td></td>
</tr>
</tbody>
</table>

Source: Hartmann (2006)
So Many Verification Measures!

- Graphical Measures
  - Scatter Plot
  - Rank Histograms
  - Box plot
  - Spread vs. Skill

- Need for easy-to-understand verification measures

- Training to understand how metrics describe different aspects of forecast quality
## RFC Verification System: Metrics

<table>
<thead>
<tr>
<th>CATEGORIES</th>
<th>DETERMINISTIC FORECAST VERIFICATION METRICS</th>
<th>PROBABILISTIC FORECAST VERIFICATION METRICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Categorical (predefined threshold, range of values)</td>
<td>Probability Of Detection (POD), False Alarm Rate (FAR), Lead Time of Detection (LTD), Critical Success Index (CSI), Pierce Skill Score (PSS), Gerrity Score (GS)</td>
<td>Brier Score (BS), Rank Probability Score (RPS)</td>
</tr>
<tr>
<td>2. Error (accuracy)</td>
<td>Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Error (ME), Bias (%), Linear Error in Probability Space (LEPS)</td>
<td>Continuous RPS</td>
</tr>
<tr>
<td>3. Correlation</td>
<td>Pearson Correlation Coefficient, Ranked correlation coefficient, scatter plots</td>
<td></td>
</tr>
<tr>
<td>4. Distribution Properties</td>
<td>Mean, variance, higher moments for observation and forecasts</td>
<td>Wilcoxon rank sum test, variance of forecasts, variance of observations, ensemble spread, Talagrand Diagram (or Rank Histogram)</td>
</tr>
</tbody>
</table>
## RFC Verification System: Metrics

<table>
<thead>
<tr>
<th>CATEGORIES</th>
<th>DETERMINISTIC FORECAST VERIFICATION METRICS</th>
<th>PROBABILISTIC FORECAST VERIFICATION METRICS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5. Skill Scores</strong>&lt;br&gt;(relative accuracy over reference forecast)</td>
<td>Root Mean Squared Error Skill Score (SS-RMSE) (with reference to persistence, climatology, lagged persistence), Wilson Score (WS), Linear Error in Probability Space Skill Score (SS-LEPS)</td>
<td>Rank Probability Skill Score, Brier Skill Score (with reference to persistence, climatology, lagged persistence)</td>
</tr>
<tr>
<td><strong>6. Conditional Statistics</strong>&lt;br&gt;(based on occurrence of specific events)</td>
<td>Relative Operating Characteristic (ROC), reliability measures, discrimination diagram, other discrimination measures</td>
<td>ROC and ROC Area, other resolution measures, reliability diagram, discrimination diagram, other discrimination measures</td>
</tr>
<tr>
<td><strong>7. Confidence</strong>&lt;br&gt;(metric uncertainty)</td>
<td>Sample size, Confidence Interval (CI)</td>
<td>Ensemble size, sample size, Confidence Interval (CI)</td>
</tr>
</tbody>
</table>
Why do we need verification measures?

• Verification statistics help in understanding
  - sources of skill in forecasts
  - sources of uncertainty in forecasts
  - conditions where and when forecasts are skillful or not skillful, and why?

• Also provide information on
  - the accuracy of forecasts
  - the improvement in terms forecast skill and decision making with alternate forecast sources (e.g., climatology, persistence, deterministic forecast)

Thus, helps in finding the limitations (flaws) of the forecast framework and, consequently helps in improving it.
Objective of diagnostic/verification tools

Assess **quality** of forecast system
i.e. determine **skill** and **value** of forecast

A forecast has **skill** if it predicts the observed conditions well according to some objective or subjective criteria.

A forecast has **value** if it helps the user to make better decisions than without knowledge of the forecast.

- Forecasts with poor skill can be valuable (e.g. extreme event forecasted in wrong place)
- Forecasts with high skill can be of little value (e.g. blue sky desert)

Source: Hagedorn (2006)
<table>
<thead>
<tr>
<th>What makes an ensemble forecast “good”?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecasts should agree with observations, with few large errors</td>
</tr>
<tr>
<td>Forecast mean should agree with observed mean</td>
</tr>
<tr>
<td>Linear relationship between forecasts and observations</td>
</tr>
<tr>
<td>Forecast should be more accurate than unskilled reference forecasts (e.g., random chance, persistence, or climatology)</td>
</tr>
</tbody>
</table>

Source: Ebert (2003)
<table>
<thead>
<tr>
<th>What makes an ensemble forecast “good”?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binned forecast values should agree with binned observations (agreement between categories)</td>
</tr>
<tr>
<td>Forecast can discriminate between events &amp; non-events</td>
</tr>
<tr>
<td>Forecast can predict extreme values</td>
</tr>
<tr>
<td>Forecast represents the associated uncertainty</td>
</tr>
</tbody>
</table>

Source: Ebert (2003)
Brier Score (BS)

**Brier Score** measures mean squared probability error

Consider a specific event by fixing a threshold, then estimate

- $p_i$, forecast probability, is fraction of members predicting event
- $o_i$, observed outcome, is “1” if event occurs, otherwise it is “0”

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$$

BS varies from 0 (perfect deterministic forecasts) to 1 (perfectly wrong)

Decomposition of BS:
BS = Reliability – Resolution + Uncertainty
Brier Skill Score (BSS)

- Brier Skill Score (BSS) measures improvement over reference

\[ BSS = 1 - \frac{BS}{BS_c} \]

Positive BSS => better than reference
Negative BSS => worse than reference

Analogous to MSE skill score
**Ranked Probability Score (RPS)**

**RPS** measures the quadratic distance between forecast and verification probabilities for several categories.

Consider multiple events by fixing multiple thresholds.

\[ p_i, \text{ forecast probability, is fraction of members predicting event} \]
\[ p_1=0.05, \ p_2=0.20, \ p_3=0.35, \ p_4=0.25, \ p_5=0.15 \]

\[ o_i, \text{ observed outcome, is “1” if event occurs, otherwise it is “0”} \]
\[ o_3=1.0, \ o_1=o_2=o_4=o_5=0.0 \]
Ranked Probability Score (RPS)

\( p_i \), forecast probability, is fraction of members predicting event

\( p_1=0.05, p_2=0.20, p_3=0.35, p_4=0.25, p_5=0.15 \)

\( o_i \), observed outcome, is “1” if event occurs, otherwise it is “0”

\( o_1=o_2=o_4=o_5=0.0, o_3=1.0 \)

\[
\text{RPS} = \frac{1}{K-1} \left[ \sum_{k=1}^{K} \left( \sum_{j=1}^{k} p_j - \sum_{j=1}^{k} o_j \right)^2 \right]
\]

\[
\text{RPS} = \frac{1}{4} \left[ (0.05-0.0)^2 + (0.25-0.0)^2 + (0.60-1.0)^2 + (0.85-1.0)^2 + (1.0-1.0)^2 \right] / 4
\]

\( \text{RPS} = 0.06 \)

If \( K=2 \), then \( \text{RPS} = \text{BS} \); Analogous to BS, but multiple categories
Ranked Probability Skill Score (RPSS)

• **RPSS** measures improvement over reference

\[
RPSS = 1 - \frac{RPS_{\text{for}}}{RPS_{\text{ref}}}
\]

RPS for forecast: \( RPS_{\text{for}} \)

RPS for reference forecast: \( RPS_{\text{ref}} \)

RPS for perfect forecast: \( RPS_{\text{per}} = 0 \)

RPSS = 0  => as good as climatology
RPSS = 1  => high skill – much better than climatology
RPSS < 0  => forecast worse than climatology
Brier Score & Ranked Probability Score

- Brier Score used for two category (yes/no) situations (e.g. T > 15°C)

- RPS takes into account ordered nature of variable ("extreme errors")

Source: Hagedorn (2006)
Deterministic Forecast Verification Measures

- **Mean error (bias)** – measures average difference between forecast and observations
  
  \[
  \text{Mean error} = \frac{1}{N} \sum_{i=1}^{N} (F_i - O_i)
  \]

- **Mean Absolute Error (MAE)** – measures average magnitude of forecast error
  
  \[
  MAE = \frac{1}{N} \sum_{i=1}^{N} |F_i - O_i|
  \]

- **Root Mean Square Error (RMSE)** – measures error magnitude, with large errors having a greater impact than in the MAE
  
  \[
  RMSE = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (F_i - O_i)^2}
  \]
Deterministic Forecast Verification Measures

- **Pearson Correlation Coefficient** – measures linear correspondence between forecasts and observations

\[
r = \frac{\sum (F - \bar{F})(O - \bar{O})}{\sqrt{\sum (F - \bar{F})^2} \sqrt{\sum (O - \bar{O})^2}}
\]
Deterministic (Yes/No) Forecast Verification Measures

**Probability of Detection (POD)** – measures fraction of events that were correctly forecast to occur

\[
POD = \frac{H}{H+M}
\]

**False Alarm Ratio (FAR)** – measures fraction of "yes" forecasts that were incorrect

\[
FAR = \frac{F}{F+H}
\]

**BIAS score** – measures ratio of forecast frequency to observed frequency

\[
BIAS = \frac{(H+F)}{(H+M)}
\]
Verification of two category (yes/no) situation

- Compute 2 x 2 contingency table: (for a set of cases)

<table>
<thead>
<tr>
<th>Event observed</th>
<th>Yes</th>
<th>No</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event forecasted</td>
<td>Yes</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>c</td>
<td>c+d</td>
</tr>
<tr>
<td>total</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d=n</td>
</tr>
</tbody>
</table>

- Event Probability: \( s = \frac{(a+c)}{n} \)
- Probability of a Forecast of occurrence: \( r = \frac{(a+b)}{n} \)
- Frequency Bias: \( B = \frac{(a+b)}{(a+c)} \)
- Hit Rate: \( H = \frac{a}{(a+c)} \)
- False Alarm Rate: \( F = \frac{b}{(b+d)} \)
- False Alarm Ratio: \( FAR = \frac{b}{(a+b)} \)

Source: Hagedorn (2006)
Example of Finley Tornado Forecasts (1884)

• Compute 2 x 2 contingency table:
  (for a set of cases)

<table>
<thead>
<tr>
<th>Event observed</th>
<th>Yes</th>
<th>No</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>28</td>
<td>72</td>
<td>100</td>
</tr>
<tr>
<td>No</td>
<td>23</td>
<td>2680</td>
<td>2703</td>
</tr>
<tr>
<td>total</td>
<td>51</td>
<td>2752</td>
<td>2803</td>
</tr>
</tbody>
</table>

• Event Probability: \( s = \frac{a+c}{n} = 0.018 \)

• Probability of a Forecast of occurrence: \( r = \frac{a+b}{n} = 0.036 \)

• Frequency Bias: \( B = \frac{a+b}{a+c} = 1.961 \)

• Hit Rate: \( H = \frac{a}{a+c} = 0.549 \)

• False Alarm Rate: \( F = \frac{b}{b+d} = 0.026 \)

• False Alarm Ratio: \( FAR = \frac{b}{a+b} = 0.720 \)

Source: Hagedorn (2006)
## Extension of 2 x 2 Contingency Table for Probabilistic Forecast

<table>
<thead>
<tr>
<th>Event forecasted</th>
<th>Event observed</th>
<th>threshold</th>
<th>$H$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;80% - 100%</td>
<td>Yes</td>
<td>30</td>
<td>&gt;80%</td>
<td>30/105</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>5</td>
<td></td>
<td>5/105</td>
</tr>
<tr>
<td>&gt;60% - 80%</td>
<td>Yes</td>
<td>25</td>
<td>&gt;60%</td>
<td>55/105</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>10</td>
<td></td>
<td>15/105</td>
</tr>
<tr>
<td>&gt;40% - 60%</td>
<td>Yes</td>
<td>20</td>
<td>&gt;40%</td>
<td>75/105</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>15</td>
<td></td>
<td>30/105</td>
</tr>
<tr>
<td>&gt;20% - 40%</td>
<td>Yes</td>
<td>15</td>
<td>&gt;20%</td>
<td>90/105</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>20</td>
<td></td>
<td>50/105</td>
</tr>
<tr>
<td>&gt;0% - 20%</td>
<td>Yes</td>
<td>10</td>
<td>&gt;0%</td>
<td>100/105</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>25</td>
<td></td>
<td>75/105</td>
</tr>
<tr>
<td>0%</td>
<td>Yes</td>
<td>5</td>
<td></td>
<td>105/105</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>30</td>
<td></td>
<td>105/105</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>105</td>
<td></td>
<td>105/105</td>
</tr>
</tbody>
</table>

Source: Hagedorn (2006)
Extension of 2 x 2 Contingency Table for Probabilistic Forecast

<table>
<thead>
<tr>
<th>Event observed</th>
<th>Yes</th>
<th>No</th>
<th>threshold</th>
<th>H</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;80% - 100%</td>
<td>30</td>
<td>5</td>
<td>&gt;80%</td>
<td>0.29</td>
<td>0.05</td>
</tr>
<tr>
<td>&gt;60% - 80%</td>
<td>25</td>
<td>10</td>
<td>&gt;60%</td>
<td>0.52</td>
<td>0.14</td>
</tr>
<tr>
<td>&gt;40% - 60%</td>
<td>20</td>
<td>15</td>
<td>&gt;40%</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>&gt;20% - 40%</td>
<td>15</td>
<td>20</td>
<td>&gt;20%</td>
<td>0.86</td>
<td>0.48</td>
</tr>
<tr>
<td>&gt;0% - 20%</td>
<td>10</td>
<td>25</td>
<td>&gt;0%</td>
<td>0.95</td>
<td>0.71</td>
</tr>
<tr>
<td>0%</td>
<td>5</td>
<td>30</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>total</td>
<td>105</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Hagedorn (2006)
Relative Operating Characteristic (ROC)

ROC measures the ability of forecast to discriminate between events and no-events

ROC curve: plot of H against F for range of probability thresholds

ROC area: area under the ROC curve; measures skill

A=0.5 => no skill
A=1 => perfect deterministic forecast

Source: Hagedorn (2006)
Relative Operating Characteristic (ROC)

Measures the ability of the forecast to discriminate between events and non-events (resolution)

→ Plot hit rate $H$ vs false alarm rate $F$ using a set of varying probability thresholds to make the yes/no decision.
  - Close to upper left corner – good resolution
  - Close to diagonal – little skill

- Area under curve ("ROC area") is a useful summary measure of forecast skill
  - Perfect: ROC area = 1
  - No skill: ROC area = 0.5
  -> ROC skill score $ROCS = 2(ROC_{area} - 0.5)$

- Not sensitive to bias.
- The ROC is conditioned on the observations (i.e., given that Y occurred, what was the corresponding forecast?)
- Reliability and ROC diagrams are good companions

Source: Ebert (2005)
Comparison of Approaches

• Brier score
  ➢ Based on squared error
  ➢ Strictly proper scoring rule
  ➢ Calibration is an important factor; lack of calibration impacts scores
  ➢ Decompositions provide insight into several performance attributes
  ➢ Dependent on frequency of occurrence of the event

• ROC
  ➢ Considers forecasts’ ability to discriminate between Yes and No events
  ➢ Calibration is not a factor
  ➢ Less dependent on frequency of occurrence of event
  ➢ Provides verification information for individual decision thresholds

Source: Barbara (2001)
Reliability Diagrams

“When you say 80% chance of high flows, how often do high flows happen?”

P(O/F)

Source: Hartmann (2006)
Graphical Representation of Measures

**Discrimination**
- Do the forecasts discriminate between types of future events?
- If a flood happened was there a forecast?
- Sort based upon the observed values

=> Discrimination diagram

\[ p(f|x=0) \text{ and } p(f|x=1) \]

**Reliability**
- When we forecast an event, are the forecasts reliable?
- If we forecast something, does it happen?
- Sort based upon the forecast values

=> Reliability diagram

\[ p(x=1|f_i) \text{ vs. } f_i \]

Forecast Reliability

*If the forecast says there’s a 50% chance of wet, wet should happen 50% of the time*

Source: Hartmann (2006)
Forecast Reliability

Reliability Diagrams identifies conditional bias

**Reliability**

P[O|F]

Does the frequency of occurrence match your probability statement?

Source: Hartmann (2006)
Reliability (Attribute) Diagram

Attributes diagram: Reliability, Resolution, Skill/No-skill

- Good reliability – close to diagonal
- Good resolution – wide range of frequency of observations corresponding to forecast probabilities
- Sharpness diagram \( p(f) \) – histogram of forecasts in each probability bin shows the sharpness of the forecast.

The reliability diagram is conditioned on the forecasts (i.e., given that \( X \) was predicted, what was the outcome?), and can be expected to give information on the real meaning of the forecast.

Source: Ebert (2005)
Reliability and Sharpness

Climatology  Minimal RESolution  Underforecasting

Good RES, at expense of REL  Reliable forecasts of rare event  Small sample size

Examples: 1) Good Reliability

Simulated flows

Postprocessed flows

Source: Seo (2005)
Examples: 1) Good Reliability

Source: Seo (2005)
Examples: 1) Good Reliability

Reliability Diagram (agreement between forecast probability and mean observed frequency)

Source: Seo (2005)
Examples: 1) Good Reliability

ROC (ability of forecast to discriminate between events & non-events)

Source: Seo (2005)
Examples: ROC

**ROC** (ability of forecast to discriminate between events & non-events)

- **Good Forecast**
- **Over estimated**
- **Random (no skill)**
Examples: 1) Good Reliability

**ROC** (ability of forecast to discriminate between events & non-events)

Source: Seo (2005)
Examples: 2) Positive Bias / Overestimated

Simulated flows

Postprocessed flows

Positive Bias

Source: Seo (2005)
Examples: 2) Positive Bias / Overestimated

Source: Seo (2005)
Examples: 2) Positive Bias / Overestimated

Reliability Diagram (agreement between forecast probability and mean observed frequency)

Source: Seo (2005)
Examples: 2) Positive Bias / Overestimated

ROC (ability of forecast to discriminate between events & non-events)

Source: Seo (2005)
Examples: 3) No Skill / Random

Simulated flows

Postprocessed flows

Source: Seo (2005)
Examples: 3) No Skill / Random

Source: Seo (2005)
Examples: 3) No Skill / Random

Reliability Diagram (agreement between forecast probability and mean observed frequency)

Source: Seo (2005)
Examples: 3) No Skill / Random

ROC (ability of forecast to discriminate between events & non-events)

Source: Seo (2005)
Discrimination Diagrams

“When dry happened, what were the forecasts up to?”

$P(F|O)$

Source: Hartmann (2006)
Discrimination Diagrams

“When dry happened, what were the forecasts up to?”

Discrimination: $P[F|O]$

Can the forecasts distinguish among different events?

**Good discrimination!**

**Not much discrimination!**

Source: Hartmann (2006)
Thank you