Seminar D: ensemble verification concepts and requirements

James Brown
james.brown@hydrosolved.com
Contents

1. Motivations for verification
2. Data requirements
3. Attributes of forecast quality
4. Measures of forecast quality
5. Final thoughts and suggestions
1. Motivations for verification
Why verify?

Forecasts incomplete if quality unknown

- Ensemble forecasts can be poor quality
- How much confidence to place in them?
- Are they unbiased and skillful? When/where/how?
- Where to focus improvements? Are they worth it?

An example: component error analysis

- Total uncertainty = meteorological + hydrological
- HEFS = MEFP + EnsPost
- Component error analysis can separate the two
Example: two very different basins

- Fort Seward, CA (FTSC1) and Dolores, CO (DOSC1)
- Total skill in EnsPost-adjusted GFS streamflow forecasts is similar
- Origins are completely different (and understandable)
**Example: two very different seasons**

- However, in FTSC1, completely different picture in wet vs. dry season
- In wet season (which dominates overall results), mainly MEFP skill
- In dry season, skill mainly originates from EnsPost (persistence)
2. Data requirements
Datasets

- Hindcasts or archived forecasts (forcing and flow)
- Reliable observations (e.g. no major ratings biases)
- Hydrologic simulations for component error analysis
- Large sample (long record) and consistent record

Verification sample size depends on

- Period of record and frequency of T0s
- Aggregation period
- Sub-setting of data (“conditional verification”)

What data are required?
How to mitigate small sample?

Steps to reduce impacts

• Hindcasting (see earlier)
• Be careful with conditioning (i.e. avoid small subsets)
• Be careful with aggregation (e.g. monthly volumes)
• Choose verification metrics that summarize quality
• Can set minimum sample size in EVS (p.104 manual)

Steps to assess impacts

• Qualitative: check sample size plots in EVS
• Quantitative: compute confidence intervals (p.48)
Data quality control (QC)

Before hindcasting: QC input data
- Use MEFP/EnsPost data and parameter diagnostics
- Check for non-physical values and outliers

After hindcasting: QC output data
- Make test runs and visualize results for gross errors
- Check all expected forecasts/members present
- Check for non-physical values and outliers
- Outliers can have a large (obscured) impact on stats
- Check verification pairs carefully…
Pairing mechanics and QC

- Pairing often requires assumptions/data manipulation
- For example, aggregation or re-timing of data
- E.g. Forecast (SQIN) vs. QME in ABRFC (GMT-6)
- **Always QC the pairs** (e.g. for 1-2 locations)!
3. Attributes of forecast quality
First, the big picture

Three separate, but related, concepts

- **Quality**: synonymous w/ verification (vs. observations)
- **Utility**: service is fit for purpose (includes quality)
- **Consistency**: forecasters not “gaming” the system

Examples of quality vs. utility

- A flood forecasting system may be reliable (quality)…
- …but forecasts may not be timely (utility)
- Climatological ensembles are unskillfull (quality)…
- …but are useful for water resources planning (utility)
Decades of publications on quality!

- Interested in forecast errors (forecast - observed)
- John Park Finley (1884): tornado verification
- Murphy and Winkler (1987): attributes of quality
- [http://hepex.irstea.fr/what-is-a-good-forecast/](http://hepex.irstea.fr/what-is-a-good-forecast/)
Two types of quality

Absolute quality vs. relative quality

- **Absolute**: properties of one system (vs. observed)
- **Relative**: comparison of two systems (vs. observed)
- **Relative quality** is also known as **skill**
- **Skill** is valuable, but choice of baseline needs thought
  - Skill (% gain) is easy to communicate, but not always to interpret
  - Think about what you want the system to improve on (e.g. EnsPost should improve on raw streamflow forecasts)
Attributes of quality

What is meant by attribute here?

• A “desirable” property of a forecasting system
• Specifically, a desirable relationship with observations
• A forecasting system has multiple attributes of quality
• Three, well-known from deterministic forecasting…

Accuracy, bias, and association

• Accuracy: generic term for total error (e.g. MSE)
• Bias: generic term for a directional error (e.g. ME)
• Association: generic for correspondence (e.g. COV)
Attributes of quality: examples

- Unbiased
- Strong association
- High accuracy (small total error)

- Some bias
- Moderate association
- Moderate accuracy (moderate total error)

- Large bias
- Strong association
- Low accuracy (high total error)

- Unbiased (but conditionally biased)
- Negative association
- Low accuracy (high total error)
Conditional attributes

Unconditional vs. conditional quality

• Unconditional
  • All data, no subsets (except by forecast lead time)

• Conditional
  • Many possible conditions; season, flow amount etc.

Let’s look at some ensemble forecasts…
Ensemble forecasts: raw data

Streamflow (Q) is both observed (Y) and forecast (X).

Consider one discrete event: exceeding a flow threshold, \( q = 5.3 \) CFS.

The forecast probability is \( f(q) = \text{prob}[X > q] \). The observed probability is \( o(q) = \text{prob}[Y > q] \).

Their “joint probability distribution” is denoted \( g(f, o) \).

<table>
<thead>
<tr>
<th>((X, Y))</th>
<th>((f(5.3), o(5.3)))</th>
</tr>
</thead>
<tbody>
<tr>
<td>({1.1, \ldots, 3.3}, \ 3.2)</td>
<td>(0.0, 0.0)</td>
</tr>
<tr>
<td>({2.6, \ldots, 21.5}, \ 20.2)</td>
<td>(0.9, 1.0)</td>
</tr>
<tr>
<td>({3.2, \ldots, 19.8}, \ 18.2)</td>
<td>(0.8, 1.0)</td>
</tr>
<tr>
<td>({4.5, \ldots, 12.5}, \ 13.4)</td>
<td>(0.7, 1.0)</td>
</tr>
<tr>
<td>({13.5, \ldots, 28.3}, \ 24.1)</td>
<td>(1.0, 1.0)</td>
</tr>
<tr>
<td>({0.2, \ldots, 7.8}, \ 2.1)</td>
<td>(0.3, 0.0)</td>
</tr>
<tr>
<td>({0.1, \ldots, 5.4}, \ 5.3)</td>
<td>(0.1, 0.0)</td>
</tr>
<tr>
<td>({7.3, \ldots, 16.5}, \ 12.4)</td>
<td>(1.0, 1.0)</td>
</tr>
<tr>
<td>({2.5, \ldots, 40.1}, \ 30.5)</td>
<td>(0.9, 1.0)</td>
</tr>
<tr>
<td>({4.9, \ldots, 57.3}, \ 47.2)</td>
<td>(0.9, 1.0)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Example of unconditional bias

The forecasts and observations should predict $Q > q$ with the same probability, on average.

\[
\text{Bias} = \frac{1}{n} \sum_{i=1}^{n} (f_i(5.3) - o_i(5.3)) \approx 0
\]

In other words:

((f(5.3), o(5.3))

(0.0, 0.0) (0.0-0.0) = 0.0
(0.9, 1.0) (0.9-1.0) = -0.1
(0.8, 1.0) (0.8-1.0) = -0.2
(0.7, 1.0) (0.7-1.0) = -0.3
(1.0, 1.0) (1.0-1.0) = 0.0
(0.3, 0.0) (0.3-0.0) = 0.3
(0.1, 0.0) (0.1-0.0) = 0.1
(1.0, 1.0) (1.0-1.0) = 0.0
(0.9, 1.0) (0.9-1.0) = -0.1
(0.9, 1.0) (0.9-1.0) = -0.1
...

Bias = -0.04
### Example of conditional bias

Given $f(5.3) = 0.9$, the forecasts are “reliable” if the event is observed 90% of the time, on average.

<table>
<thead>
<tr>
<th>$(f(5.3), o(5.3))$</th>
<th>$(f(5.3) - o(5.3))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(0.0, 0.0)$</td>
<td>$(0.0-0.0) = 0.0$</td>
</tr>
<tr>
<td>$(0.9, 1.0)$</td>
<td>$(0.9 - 1.0) = -0.1$</td>
</tr>
<tr>
<td>$(0.8, 1.0)$</td>
<td>$(0.8 - 1.0) = -0.2$</td>
</tr>
<tr>
<td>$(0.7, 1.0)$</td>
<td>$(0.7 - 1.0) = -0.3$</td>
</tr>
<tr>
<td>$(1.0, 1.0)$</td>
<td>$(1.0 - 1.0) = 0.0$</td>
</tr>
<tr>
<td>$(0.3, 0.0)$</td>
<td>$(0.3 - 0.0) = 0.3$</td>
</tr>
<tr>
<td>$(0.1, 0.0)$</td>
<td>$(0.1 - 0.0) = 0.1$</td>
</tr>
<tr>
<td>$(1.0, 1.0)$</td>
<td>$(1.0 - 1.0) = 0.0$</td>
</tr>
<tr>
<td>$(0.9, 1.0)$</td>
<td>$(0.9 - 1.0) = -0.1$</td>
</tr>
<tr>
<td>$(0.9, 1.0)$</td>
<td>$(0.9 - 1.0) = -0.1$</td>
</tr>
<tr>
<td>$(0.9, 1.0)$</td>
<td>$(0.9 - 1.0) = -0.1$</td>
</tr>
</tbody>
</table>

In other words:

$$\sum_{f(5.3) = 0.9} \left(0.9 - o(5.3)\right) \approx 0$$

**Bias** = $-0.1$

In practice, $n>>3$ is needed!
Attributes of probability forecasts

\[ g(f,o) = r(o|f)s(f) \]  \hspace{1cm} “Calibration-refinement”

\[ g(f,o) = v(f|o)u(o) \]  \hspace{1cm} “Likelihood-base-rate”

“Sharpness” is concerned with \( s(f) \)

“Uncertainty” is concerned with \( u(o) \)

“Reliability” is concerned with \( r(o|f) \) vs. \( s(f) \)

“Resolution” is concerned with \( r(o|f) \)

“Discrimination” is concerned with \( v(f|o) \)

“Type-II bias” is concerned with \( v(f|o) \) vs. \( u(o) \)
4. Measures of forecast quality
Tips on selecting measures

Things to consider

- The study may address specific users/applications
- But, do not rely on any single measure of quality
- Build a picture across several attributes of quality
  - Overall impression of accuracy (total error)
  - Unconditional and conditional biases (directional error)
  - Measures that are insensitive to bias (correlation, discrimination)
  - Skill relative to a baseline (remember skill reflects the baseline!)
- Be mindful of sample size issues
- Extreme events: be mindful of non-occurrences!
Extreme events: tornado forecasts

John Park Finley: 1854-1943

<table>
<thead>
<tr>
<th>Observed</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>28</td>
</tr>
<tr>
<td>No</td>
<td>23</td>
</tr>
</tbody>
</table>

Correct:
\[
\frac{28 + 2680}{28 + 72 + 23 + 2680} = 96.5\%
\]

Correct if always forecasting “no tornado”:
\[
\frac{72 + 2680}{28 + 72 + 23 + 2680} = 98.1\%!
\]

Correct when tornado observed:
\[
\frac{28}{28 + 72} = 28\%
\]
### What measures in EVS?

<table>
<thead>
<tr>
<th>Metric name</th>
<th>Feature tested</th>
<th>Discrete events?</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean error</td>
<td>Ensemble average</td>
<td>No</td>
<td>Lowest</td>
</tr>
<tr>
<td>Relative mean error</td>
<td>Ensemble average</td>
<td>No</td>
<td>Lowest</td>
</tr>
<tr>
<td>RMSE</td>
<td>Ensemble average</td>
<td>No</td>
<td>Lowest</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>Ensemble average</td>
<td>No</td>
<td>Lowest</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>Ensemble average</td>
<td>No</td>
<td>Lowest</td>
</tr>
<tr>
<td>Brier Score</td>
<td>Lumped error score</td>
<td>Yes</td>
<td>Low</td>
</tr>
<tr>
<td>Mean CRPS</td>
<td>Lumped error score</td>
<td>No</td>
<td>Low</td>
</tr>
<tr>
<td>Mean error in prob.</td>
<td>Reliability (unconditional bias)</td>
<td>No</td>
<td>Low</td>
</tr>
<tr>
<td>Brier Skill Score</td>
<td>Lumped error score vs. reference</td>
<td>Yes</td>
<td>Low</td>
</tr>
<tr>
<td>ROC score</td>
<td>Lumped discrimination score</td>
<td>Yes</td>
<td>Low</td>
</tr>
<tr>
<td>Mean CRPSS</td>
<td>Lumped error score vs. reference</td>
<td>No</td>
<td>Low</td>
</tr>
<tr>
<td>Spread-bias diagram</td>
<td>Reliability (conditional bias)</td>
<td>No</td>
<td>High</td>
</tr>
<tr>
<td>Rank histogram</td>
<td>Reliability (conditional bias)</td>
<td>No</td>
<td>High</td>
</tr>
<tr>
<td>Reliability diagram</td>
<td>Reliability (conditional bias)</td>
<td>Yes</td>
<td>High</td>
</tr>
<tr>
<td>ROC diagram</td>
<td>Discrimination</td>
<td>Yes</td>
<td>High</td>
</tr>
<tr>
<td>Modified box plots</td>
<td>Error visualization</td>
<td>No</td>
<td>Highest</td>
</tr>
</tbody>
</table>
Accuracy (total error): mean CRPS

CRPS = \int_{-\infty}^{\infty} (f_i(q) - o_i(q))^2 dq

- Then average across multiple forecasts
- Small scores = better
- Skill score “% gain”:

\[
CRPSS = 1 - \frac{CRPS_{\text{MAIN}}}{CRPS_{\text{REFERENCE}}}
\]
Accurate (total error): Brier Score

**Observed:**
\[ o_i(q) = \text{Prob}[Y \leq q] \]

**Forecast:**
\[ f_i(q) = \text{Prob}[X \leq q] \]

- BS for a discrete flow threshold, \( q = 5.3 \)
  \[ BS = \frac{1}{n} \sum_{i=1}^{n} [f_i(5.3) - o_i(5.3)]^2 \]
- Mean square error in probability over \( n \) pairs
- Small scores = better
- Skill score available

**Flow (Q) [cfs]:**

**Cumulative probability:**

**Error:**

- Observed: \( o_i(5.3) = 0.0 \)
- Forecast: \( f_i(5.3) = 0.21 \)
Looks at discrete forecast, i.e. one event only (e.g. flooding).

“When flooding is forecast with probability 0.48, it should occur 48% of the time.” Actually occurs 36% of time.

Flooding forecast 23 times with probability 0.4-0.6 (mean=0.48)
Conditional bias: box plots

‘Error’ for 1 forecast
- Largest +ve error
- 90 percent
- 80 percent
- Median error
- 20 percent.
- 10 percent.
- Largest –ve error

MEFP precipitation ensembles (1 day ahead total)

Zero error line

“Blown forecasts”

Precipitation is bounded at 0

A ‘Type-II conditional bias’, i.e. depends on observed
Discrimination: ROC

- **Probability of Detection** \(\frac{TP}{TP+FN}\)
- **Probability of False Detection** \(\frac{FP}{FP+TN}\)

**Legend**:
- **flood**
  - TP (True Positive)
  - FP (False Positive)
- **!flood**
  - FN (False Negative)
  - TN (True Negative)

**Climatological prob. forecast**
- "sitting on the fence"

**Warn flood (W) when y>0.1**
- "OK to cry wolf!"

**Warn flood (W) when y>0.9**
- "Must not cry wolf!"

- **Perfect**
- **Looks at discrete forecast, i.e. one event only (e.g. flooding).**
5. Final thoughts and suggestions
Final thoughts

Things to consider

• Try to maximize period and consistency of record
• Due diligence before verification (data/calibration QC)
• Always QC the paired data, as mistakes easily made
• Identify the scope/users of the verification (questions)
• Consider several attributes and measures of quality
• Consider contrasting attributes (e.g. bias/association)
• Be mindful of sample sizes and verify accordingly
• Don’t be afraid to explore results iteratively!
Resources and references


Extra slides
1. What do I want to know?

How reliable were spring flood ESPs in NCRFC from 1980-2010?

2. What data and how to subset? flow > flood && ‘spring’

3. Produce and QC raw data (pairs)

4. What measures of quality?

5. Interpret measures: do they answer the questions?
Structured user interface

1. Verification (per location)
   - Specify locations, data sources, metrics etc.

2. Aggregation (many locations): option
   - Choose locations, aggregation method etc.

3. Output (graphical and numerical)
Data QC example

- Cannonsville, NY (CNNN6): reservoir inflows are estimated
- Inflow estimates do not include evaporation = biases in dry conditions
- Data QC problems can be insidious (e.g. masked by model errors)
Pairing tips

Things to remember when pairing

- Forecasts/simulations in UTC (12Z, $\Delta t=1$ or 6 hours)
- Observations in local time (e.g. 5Z, 11Z,.. in MARFC)
- Observations generally enforced as CST for pairing…
  - …avoids interpolation, but adds error for non-CST
  - …except where forecasts are hourly (then, no error)
- Remember, wrong pairs can be created quite easily…
  - …especially when forecasts are hourly (CB, CN)
- So, always QC the pairs (see exercises)!
**Unconditional bias: MEPD**

- Recall example of Cannonsville, NY (CNNN6) with dry bias
- Mean Error of Probability Diagram: average forecast CDF vs. observed
- Shows climatological bias in the forecasts, i.e. mean probability error

\[
\bar{f}_i(q) = \frac{1}{n} \sum_{i=1}^{n} f_i(q) \quad \forall q
\]

\[
\bar{o}_i(q) = \frac{1}{n} \sum_{i=1}^{n} o_i(q) \quad \forall q
\]

Unbiased: \( E[f(q) - o(q)] = 0 \)