Ensemble Verification Metrics

Debbie Hudson (Bureau of Meteorology, Australia)

ECMWF Annual Seminar 2017

Acknowledgements: Beth Ebert
Overview

1. Introduction
2. Attributes of forecast quality
3. Metrics: full ensemble
4. Metrics: probabilistic forecasts
5. Metrics: ensemble mean
6. Key considerations: sampling issues; stratification; uncertainty; communicating verification
Purposes of ensemble verification

User-oriented
- How accurate are the forecasts?
- Do they enable better decisions than could be made using alternate information (persistence, climatology)?

Inter-comparison and monitoring
- How do forecast systems differ in performance?
- How does performance change over time?

Calibration
- Assist in bias removal and downscaling

Diagnosis
- Pinpoint sources of error in ensemble forecast system
- Diagnose impact of model improvements, changes to DA and/or ensemble generation etc.
- Diagnose/understand mechanisms and sources of predictability
Evaluating Forecast Quality

Need large number of forecasts and observations to evaluate ensembles and probability forecasts

Forecast quality vs. value

Attributes of forecast quality:

- Accuracy
- Skill
- Reliability
- Discrimination and resolution
- Sharpness
Accuracy and Skill

Accuracy

Overall correspondence/level of agreement between forecasts and observations

Skill

A set of forecasts is skilful if better than a reference set, i.e. skill is a comparative quantity

Reference set e.g., persistence, climatology, random

\[
Skill \ Score = \frac{score_{\text{forecast}} - score_{\text{reference}}}{score_{\text{perfect forecast}} - score_{\text{reference}}}
\]
Reliability

Ability to give unbiased probability estimates for dichotomous (yes/no) forecasts

Can I trust the probabilities?

Defines whether the certainty communicated in the forecasts is appropriate

Forecast distribution represents distribution of observations

Reliability can be improved by calibration
Resolution

• How much does the observed outcome change as the forecast changes i.e., "Do outcomes differ given different forecasts?"
• Conditioned on the forecasts

Discrimination

• Can different observed outcomes can be discriminated by the forecasts.
• Conditioned on the observations

Indicates potential "usefulness"
Cannot be improved by calibration
Discrimination

(a) Good discrimination
(b) Poor discrimination
(c) Good discrimination
Sharpness is tendency to forecast extreme values (probabilities near 0 or 100%) rather than values clustered around the mean (a forecast of climatology has no sharpness).

A property of the forecast only.

Sharp forecasts are "useful" BUT don’t want sharp forecasts if not reliable. Implies unrealistic confidence.
What are we verifying?

How are the forecasts being used?

Ensemble distribution

Set of forecasts making up the ensemble distribution

Use individual members or fit distribution

Probabilistic forecasts generated from the ensemble

Create probabilities by applying thresholds

Ensemble mean
Commonly used verification metrics

Characteristics of the full ensemble

- Rank histogram
- Spread vs. skill
- Continuous Ranked Probability Score (CRPS) (discussed under probability forecasts)
Rank histogram

Measures consistency and reliability: the observation is statistically indistinguishable from the ensemble members

→ For each observation, rank the N ensemble members from lowest to highest and identify rank of observation with respect to the forecasts

Example for 10 ensemble members

Need lots of samples to evaluate the ensemble
Rank histogram

Negative bias

Positive bias

Consistent/Reliable

Under-dispersive (overconfident)

Over-dispersive (underconfident)
Flat rank histogram does not necessarily indicate a skillful forecast.

Rank histogram shows conditional/unconditional biases BUT not full picture

- Only measures whether the observed probability distribution is well represented by the ensemble.
- Does NOT show sharpness – climatological forecasts are perfectly consistent (flat rank histogram) but not useful
Spread-skill evaluation

500 hPa Geopotential Height (20-60S)

Underdispersed (overconfident)

\[ S_{\text{ens}} < \text{RMSE} \]

Seasonal prediction system where ensemble is generated using:
A) Stochastic physics only
Spread-skill evaluation

500 hPa Geopotential Height (20-60S)

Underdispersed (overconfident)
\[ S_{\text{ens}} < \text{RMSE} \]

Consistent/reliable
\[ S_{\text{ens}} \approx \text{RMSE} \]

Overdispersed (underconfident)
\[ S_{\text{ens}} > \text{RMSE} \]

Seasonal prediction system where ensemble is generated using:
A) Stochastic physics only
B) Stochastic physics AND perturbed initial conditions

Hudson et al (2017)
Commonly used verification metrics

Probability forecasts

- Reliability/Attributes diagram
- Brier Score (BS and BSS)
- Ranked Probability Score (RPS and RPSS)
- Continuous Ranked Probability Score (CRPS and CRPSS)
- Relative Operating Characteristic (ROC and ROCS)
- Generalized Discrimination Score (GDS)
Dichotomous forecasts

Measures how well the predicted probabilities of an event correspond to their observed frequencies (reliability)

→ Plot observed frequency against forecast probability for all probability categories
→ Need a big enough sample

Histogram: how often each probability was issued.

Shows sharpness and potential sampling issues

Curve tells what the observed frequency was for a given forecast probability.
Conditioned on the forecasts
Interpretation of reliability diagrams

Reliability diagram: Example

Predictions of above normal seasonal SON rainfall

Statistical forecast scheme

Dynamical forecast scheme

Most of the forecasts issued have probabilities near 50%

Size of the circles are proportional to the number of forecasts issuing that probability

The statistical system often gave forecasts close to climatology – reliable BUT poor sharpness. Of limited use for decision-makers!
Brier score (BS)

Dichotomous forecasts
Brier score measures the mean squared probability error

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

- Score range: 0 to 1; Perfect BS: 0

Murphy's (1973) decomposition into 3 terms (for \( K \) probability classes and \( N \) samples):

\[
BS = \frac{1}{N} \sum_{k=1}^{K} n_k (p_k - \bar{o}_k)^2 - \frac{1}{N} \sum_{k=1}^{K} n_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})
\]

- Usefulness for exploring dependence of probability forecasts on ensemble characteristics
- Uncertainty term measures the variability of the observations. Has nothing to do with forecast quality!
- BS is sensitive to the climatological frequency of an event: the more rare an event, the easier it is to get a good BS without having any real skill
BS, Brier Skill Score (BSS) and the Attributes diagram

Reliability term (BS_{rel}): measures deviation of the curve from the diagonal line – error in the probabilities.

Resolution term (BS_{res}): measures deviation of the curve from the sample climate horizontal line – indicates degree to which forecast can separate different situations.

Brier skill score: measures the relative skill of the forecast compared to climatology

\[
BSS = 1 - \frac{BS}{BS_{clim}}
\]

Perfect: \(BSS = 1.0\)
Climatology: \(BSS = 0.0\)

Points in shaded region contribute to positive BSS
**BS_{rel} and BS_{res}: Example**

Aug-Sep-Oct season

**ACCESS-S1**

Probability seasonal mean rainfall above-normal over Australia

**POAMA**

Reliability (BS_{rel})

- Smaller is better

Resolution (BS_{res})

- Bigger is better
Continuous ranked probability score (CRPS) measures the difference between the forecast and observed CDFs.

\[
CRPS = \int_{-\infty}^{\infty} (P_{\text{fcst}}(x) - P_{\text{obs}}(x))^2 dx
\]

- Same as Brier score integrated over all thresholds
- On continuous scale: does not need reduction of ensemble forecasts to discrete probabilities of binary or categorical events (for multi-category use Ranked Probability Score)
- Same as Mean Absolute Error for deterministic forecasts
- Has dimensions of observed variable
- Perfect score: 0
- Rewards small spread (sharpness) if the forecast is accurate
- Skill score wrt climatology: \( CRPSS = 1 - \frac{CRPS}{CRPS_{\text{clim}}} \)
Relative Operating Characteristic (ROC)

Dichotomous forecasts
Measures the ability of the forecast to discriminate between events and non-events (discrimination)

→ Plot hit rate vs false alarm rate using a set of varying probability thresholds to make the yes/no decision.

Close to upper left corner – good discrimination
Close to or below diagonal – poor discrimination
Relative Operating Characteristic (ROC)

Dichotomous forecasts
Measures the ability of the forecast to discriminate between events and non-events (discrimination)

→ Plot hit rate vs false alarm rate using a set of varying probability thresholds to make the yes/no decision.

  Close to upper left corner – good discrimination
  Close to or below diagonal – poor discrimination

• Area under curve ("ROC area") is a useful summary measure of forecast skill
**Relative Operating Characteristic (ROC)**

Dichotomous forecasts
Measures the ability of the forecast to discriminate between events and non-events (discrimination)

→ Plot hit rate vs false alarm rate using a set of varying probability thresholds to make the yes/no decision.

- Close to upper left corner – good discrimination
- Close to or below diagonal – poor discrimination

- Area under curve ("ROC area") is a useful summary measure of forecast skill

![ROC Area Diagram](image)
Relative Operating Characteristic (ROC)

Dichotomous forecasts
Measures the ability of the forecast to discriminate between events and non-events (discrimination)

→ Plot hit rate vs false alarm rate using a set of varying probability thresholds to make the yes/no decision.

- Close to upper left corner – good discrimination
- Close to or below diagonal – poor discrimination

- Area under curve ("ROC area") is a useful summary measure of forecast skill

- ROC skill score: $\text{ROCS} = 2(\text{ROC area} - 0.5)$
- The ROC is conditioned on the observations
- Reliability and ROC diagrams are good companions
ROC: Example

ROC area of probability of a heatwave for all forecasts initialised in DJF

Hudson and Marshall (2016)
Generalized Discrimination Score (GDS)

Binary, multi-category & continuous
Rank-based measure of discrimination - does the forecast successfully rank (discriminate) the two different observations?

GDS equivalent to ROC area for dichotomous forecasts & has same scaling

\[
\text{GDS} = \frac{\text{number of successful rankings}}{\frac{1}{2} \times \text{total rankings}}
\]

Mason & Weigel (2009); Weigel & Mason (2011)
GDS (and ROC): Example

Forecast of seasonal SON rainfall

Generalized discrimination score
Seasonal (SON) precipitation from ECMWF SYSTEM4 forecasts
initialised in August verified against ERA-INT for 1981-2014

No/Poor Discrimination  Good discrimination

https://meteoswiss-climate.shinyapps.io/skill_metrics/
Commonly used verification metrics

Ensemble mean

e.g., RMSE, correlation
Debate as to whether or not this is a good idea:

**Pros:**
- Ensemble mean filters out smaller unpredictable scales
- Needed for spread – skill evaluation
- Forecasters and others use ensemble mean

**Cons:**
- Not a realization of the ensemble
- Different statistical properties to ensemble and observations

**Scores:**
- RMSE
- Anomaly correlation
- Other deterministic verification scores
Key considerations: Sampling issues

Rare and extreme events

See Chris Ferro's talk on verification of extremes

Difficult to verify probabilities on the "tail" of the PDF

- Too few samples to get robust statistics, especially for reliability
- Finite number of ensemble members may not resolve tail of forecast PDF

Size of ensemble vs number of verification samples

Robustness of verification depends on both!!!
Key considerations: Stratification

Verification results vary with region, season, climate driver……

Pooling samples can mask variations in forecast performance

Stratify data into sub-samples

• BUT must have enough samples to give robust statistics!

Example:
MJO Bivariate correlation for RMM index

Hudson et al (2017)
Key considerations: Uncertainty

Are the forecasts significantly better than a reference forecast?
Does ensemble A perform significantly better than ensemble B?

- Take into account sampling variability
- Significance levels and/or confidence intervals
- Non-parametric resampling methods (Monte Carlo, bootstrap)

Effects of observation errors

- Adds uncertainty to verification results
- True forecast skill unknown
- Extra dispersion of observed PDF
- Active area of research
Key considerations: Communicating verification to users

- Challenging to communicate ensemble verification
- Forecast quality does not necessarily reflect value
- Summary skill measure – average skill over reforecasts. Does not show how skill changes over time (windows of forecast opportunity)
- Large sampling uncertainty around scores for quantities that are of most interest to the user e.g. regional rainfall

Related considerations:

- Using reforecasts to estimate skill (smaller ensemble size that real-time)
- Models becoming more computationally expensive – constraints on reforecast size. What is optimal reforecast size – # years; start dates and ensemble size? (from a sub-seasonal to seasonal forecasting perspective)
Thanks Ian Jolliffe and Beth Ebert
Useful general references

WMO Verification working group forecast verification web page:


Special issues of *Meteorological Applications* on Forecast Verification (Vol 15 2008 & Vol 20 2013)

Thank you…

Debbie Hudson
Debbie.Hudson@bom.gov.au