Glossary of Forecast Verification Terms

David B. Stephenson

Many common words and phrases have developed very specific meanings in forecast verification studies. This glossary aims to provide clear explanations and consistent mathematical definitions for the more commonly used expressions. The population mean of a quantity B over all possible values of A for cases where condition C holds true is denoted by $E_A(B|C)$ (the conditional expectation of B over all values of A conditioned on C). The population variance of a quantity B over all values of A is denoted by $\mathrm{var}_A(B)$. Sample estimates of these population quantities are obtained by calculating the sample mean and variance over the appropriate subsample of cases where the condition is valid (stratified/composite means and variances). More comprehensive descriptions of the following can be found in indexed entries discussed in preceding chapters of this book.

artificial skill.

An overestimate of the real skill of a forecasting system caused by including the same data to evaluate the forecast skill as was used to develop/train the forecasting system. Artificial skill can be avoided by using independent training and assessment data sets. Artificial skill often occurs in practice due to the presence of long-term trends in the data set.

attributes.

Forecast quality is a multidimensional concept described by several different scalar attributes such as **overall bias**, **reliability**/**calibration (**Type 1 **conditional bias)**, **uncertainty**, **sharpness/refinement**, **accuracy**, **association**, **resolution**, and **discrimination**. All of these attributes provide useful information about the performance of a forecasting system – no single measure is sufficient for judging and comparing forecast quality.

accuracy.

The average distance/error between forecasts and observations that depends on **bias**, **resolution**, and **uncertainty** attributes. Often estimated using **Mean Squared Error** but can be estimated more robustly using statistics such as **Mean Absolute Error** that are less sensitive (more resistant) to large outlier errors.

association.

Overall strength of the relationship/dependency between the forecasts and observations that is independent of the marginal distributions. Linear association is often estimated using the product moment **correlation coefficient**.

base rate.

The marginal probability distribution, p(x), of the observations. In other words, the **sample climatology** of the event independent of any forecasts.

bias.

The difference between the central locations of the forecasts and the observations (also known as **overall bias**, **systematic bias**, or **unconditional bias)**. Most easily quantified using the **Mean Error**, $E(\hat{X})-E(X)$, i.e. the difference between the means of the forecasts and the

observations. For **categorical forecasts**, bias in marginal probabilities is estimated by the ratio of the total number of events forecast to the total number of events observed (i.e. $(a+b)/(a+c)$ for binary categorical forecasts - see **contingency table**).

Brier probability score.

The Brier score is the mean square error of probability forecasts for a binary event $X = 0,1$. It is defined as $B = E[(\hat{p} - X)^2]$ and is zero for a perfect (deterministic) forecasts and equals 1 for forecasts that are always incorrect.

calibration.

See **reliability**.

categorical forecast.

A forecast in which a discrete number of K categories of events are forecast. Categories can be either **nominal** (no natural ordering – e.g. clear, cloudy, rain) or **ordinal** (the order matters – e.g. cold, normal, warm). Categorical forecasts can be either **deterministic** (a particular category e.g. rain or no-rain tomorrow) or **probabilistic** (probabilities for each category e.g. probability of 0.3 of rain and 0.7 for no-rain tomorrow).

conditional bias

Conditional bias is the difference $E_A(A|B) - B$ between the conditional mean $E_A(A|B)$ (the average over all possible values of A for a given value of B) of a random variable A and the conditioning variable B. The conditional bias is zero when the linear regression of A on B has a slope equal to one and an intercept of zero. Type 1 conditional bias $E_X(X|\hat{X}) - \hat{X}$ is obtained by calculating the mean of the observations for particular values of the forecast, whereas type 2 conditional bias $E_{\hat{X}}(\hat{X}|X) - X$ is obtained by conditioning on the observed values. Meaures of overall conditional bias can be obtained by averaging the mean squared bias over all possible values of the conditioning variable e.g. $E_{\hat{X}}[(E_X(X|\hat{X}) - \hat{X})^2]$. Type 1 conditional bias is also known as **reliability** or **calibration** and is 0 for all values of $\hat{X}$ for a *perfectly reliable (well-calibrated)* forecasting system. A **reliability diagram** can be made by plotting $E_X(X|\hat{X})$ against $\hat{X}$.

conditional distribution.

The probability distribution of a variable, given that a related variable is restricted to a certain value. The conditional distribution of the forecasts given the observations, $p(\hat{x}|x)$, determines the **discrimination** or **likelihood**. The conditional distribution of the observations given the forecasts, $p(x|\hat{x})$, determines the **calibration** or **reliability**. These two conditional distributions are related to each via Bayes' theorem $p(x|\hat{x})p(\hat{x}) = p(\hat{x}|x)p(x)$.

contingency table.

A two-way contingency table is a two-dimensional table that gives the discrete joint sample distribution of forecasts and observations in terms of cell counts. For dichotomous categorical forecasts, having only two possible outcomes (Yes or No), the following (2x2) contingency table can be defined:

		Event Observed: Yes	Event Observed: No	Total forecast
Event Forecast	Yes	a (hits)	b (false alarms)	a+b
	No	c (misses)	d (correct rejections)	c+d
	Total observed	a+c	b+d	a+b+c+d=n

Cell count **a** is the number of event forecasts that correspond to event observations, or the number of **hits**; cell count **b** is the number of event forecasts that do not correspond to observed events, or the number of **false alarms**; cell count **c** is the number of no-event forecasts corresponding to observed events, or the number of **misses**; and cell count **d** is the number of no-event forecasts corresponding to no events observed, or the number of **correct rejections**.

Forecast quality for this (2x2) binary situation can be assessed using a surprisingly large number of different measures e.g. **percent correct** (PC), **probability of detection** (POD), **false alarm ratio** (FAR), **success ratio** (SR), **threat score** (TS) or **critical success index** (CSI), **Heidke skill score** (HSS), and a categorical measure of **bias**, etc.

correct rejection.

In a categorical verification problem, a no-event forecast that is associated with no event observed. See **contingency table**.

correlation coefficient.

A measure of the **association** between the forecasts and observations independent of the mean and variance of the marginal distributions. The Pearson product moment correlation coefficient is a measure of linear association and is invariant under any shifts or rescalings of the forecast or observed variables. The Spearman rank correlation coefficient measures monotonicity in the relationship and is invariant under any monotonic transformations of either the forecast or observed variables.

critical success index (CSI).

Also called the **threat score** (TS) and the **Gilbert score (GS)** , the CSI is a verification measure of categorical forecast performance equal to a/(a+b+c) i.e. the total number of correct event forecasts (hits) divided by the total number of event forecasts plus the number of misses (hits + false alarms + misses). The CSI is not affected by the number of non-event forecasts that are not observed (correct rejections) and is therefore strongly dependent upon the **base rate**.

deterministic forecasts.

Nonprobabilistic forecasts of either a specific category or particular value for either a discrete or continous variable. Deterministic forecasts of continuous variables are also known as **point forecasts**. Deterministic forecasts fail to provide any estimates of possible uncertainty, and this leads to less optimal decision-making than can be obtained using **probabilistic forecasts.** Deterministic forecasts are often interpreted as probabilistic forecasts having only probabilities of 0 and 1 (i.e. no uncertainty), yet it is more realistic to interpret them as probabilistic forecasts in which the uncertainty is not provided (i.e. unknown uncertainty). Sometimes (confusingly) referred to as categorical forecasts in the earlier literature.

discrimination.

The sensitivity of the likelihood $p(\hat{x} \mid x)$ to different observed values of x. It can be measured for a particular forecast value $\hat{x}$ by the **likelihood-ratio** $p(\hat{x} \mid x_1) / p(\hat{x} \mid x_2)$. A single overall summary measure is provided by the variance $\mathrm{var}_X[E_{\hat{X}}(\hat{X} \mid X)]$ of the means of the forecasts conditioned (stratified) on the observations.

equitable/equitability.

A metaverification property for screening of suitable scores for deterministic categorical forecasts (Gandin and Murphy 1992). An equitable score takes the same, no-skill value for

random forecasts and for unvarying forecasts of a constant category . This criterion is based on the principle that random forecasts or constant forecasts of a category should have the same expected no-skill score (Murphy and Daan 1985).

equitable threat score (ETS).

This score is commonly used for the verification of deterministic forecasts of rare events (e.g. precipitation amounts above a large threshold). It was developed by Gilbert (1884) as a modification of the **threat score** to allow for the number of hits that would have been obtained purely by chance – see Schaefer (1990) and Doswell et al. (1990). It is sometimes referred to as Gilbert's Skill Score. In terms of raw cell counts it is defined as

$$\frac{a-a_r}{a-a_r+b+c}$$

where $a_r=(a+b)(a+c)/n$ is the number of hits expected for forecasts independent of observations (pure chance). Note that the appearance of n in the expression for a_r means that the equitable threat score (unlike the threat score) depends explicitly on the number of correct rejections, d.

false alarm.

In a categorical verification problem, an event forecast that is associated with no event observed. See **contingency table**.

false alarm rate (F).

A verification measure of categorical forecast performance equal to the number of false alarms divided by the total number of events observed. For the (2x2) verification problem in the definition of **contingency table**, $F=b/(b+d)$. Not to be confused with **false alarm ratio**.

false alarm ratio (FAR).

A verification measure of categorical forecast performance equal to the number of false alarms divided by the total number of event forecast. For the (2x2) verification problem in the definition of **contingency table**, FAR= b/(a+b). Not to be confused with **false alarm *rate*** that is conditioned on observations rather than forecasts.

forecast verification.

The process of summarising and assessing the overall **forecast quality** of previous sets of forecasts. Although more commonly referred to as *forecast evaluation* in other disciplines, in the meteorological context forecast evaluation implies the study of user-specific **forecast value** rather than **forecast quality**. Philosophically, the word *verification* is a misnomer since all forecasts eventually fail and so can only be *falsified* not *verified*.

forecast value

The economic utility of forecasts for a particular set of forecast users often based on simple cost-loss models. Often strongly dependent on the marginal distributions and forecast bias due to users incurring very different losses for different categories of events.

forecast quality

Statistical description of how well the forecasts match the observations that provides important feedback on the forecasting system. Unlike **forecast value,** it aims to provide an overall summary of the agreement between forecasts and observations that does not depend on a

particular user's requirements. Forecast quality has many different **attributes** that can all provide useful information on the performance.

Gilbert score (GS).

Same as **critical success index** (CSI).

Gilbert's Skill Score (GSS).

Same as **equitable threat score** (ETS).

Heidke skill score (HSS).

A skill score of categorical forecast performance based on the **proportion correct (PC)** that takes into account the number of hits due to chance. Hits due to chance is given as the event relative frequency multiplied by the number of event forecasts.

hit.

A forecasted categorical event that is later observed to happen. See **contingency table**.

hit rate (H).

A categorical forecast score equal to the total number of correct event forecasts (hits) divided by the total number of events observed i.e. a/(a+c) in the (2x2) contingency table. Also known as the **probability of detection (POD)** in the older literature.

joint distribution.

The probability distribution defined over two or more variables. For independent events (i.e. no serial or spatial dependency), the joint distribution of the forecasts and observations, $p(\hat{x}, x)$, contains all of the probabilistic information relevant to the verification problem. The joint distribution can be factored into **conditional distributions** and **marginal distributions** in either of two ways:

- The **calibration-refinement** factorization $p(\hat{x}, x) = p(x \mid \hat{x}) p(\hat{x})$
- The **likelihood-base rate** factorization $p(\hat{x}, x) = p(\hat{x} \mid x) p(x)$

See Murphy and Winkler (1987) for an elegant exposition on this general and powerful framework.

likelihood.

The probability $p(\hat{x} \mid x)$ of a forecast value given a particular observed value. The sensitivity of the likelihood to the observed value determines the **discrimination** of the system. Note that the concept of likelihood is fundamental in much of statistical inference, but the usage of the term in that context is somewhat different.

marginal distribution

The probability distribution of a single variable e.g. $p(x)$ or $p(\hat{x})$. The marginal distribution of the observations, $p(x)$, is referred to as the **base rate**. See also **uncertainty**, **sharpness** and **refinement**.

Mean Absolute Error (MAE).

The mean of the absolute differences between the forecasts and observations $E(|\hat{X} - X|)$. A more robust measure of forecast accuracy than **Mean Squared Error** that is somewhat more resistant to the presence of large outlier errors. Can be made dimensionless and more stable by dividing by the mean absolute deviation of the observations $E(|X - E(X)|)$ to yield a **Relative Absolute Error**.

Mean Error (ME).

The mean of the differences of the forecasts and observations $E(\hat{X} - X) = E(\hat{X}) - E(X)$. It is an overall measure of the unconditional bias of the forecasts (see **reliability**).

Mean Square Error (MSE).

The mean of the squares of the differences of the forecasts and observations $E[(\hat{X} - X)^2]$. It is a widely-used measure of forecast **accuracy** that depends on **bias**, **resolution**, and **uncertainty**. Because it is a quadratic loss function, it can be overly sensitive to large outlier forecast errors and is therefore a non-resistant measure (see **Mean Absolute Error**). The MSE can sometimes encourage forecasters to hedge towards forecasting smaller than observed variations in order to reduce the risk of making a large error.

metaverification.

The screening of suitable scores by requiring desirable properties such as **propriety**, **equitability**, etc.

miss.

See **contingency table**.

non-probabilistic forecast.

See **deterministic forecast**.

percent correct (PC).

The percentage of correct categorical forecasts (hits and correct rejections) equal to (a+d)/(a+b+c+d) for the (2x2) problem (see **contingency table**).

predictand.

The observable object x that is to be forecast. In regression, the predictand is known as the **response variable**, which is predicted using the **predictor**. Scalar predictands can be nominal categories (e.g. snow, foggy, sunny), ordinal categories (e.g. cold, normal, hot), discrete variables (e.g. number of hurricanes), or continuous variables (e.g. temperature).

predictor.

A forecast of either the value $\hat{x}$ of a predictand (deterministic forecasts) or the probability distribution $\hat{p}(x)$ of a predictand (probabilistic forecasts). In regression, the predictor(s) is the predicted value of the **predictand** calculated using knowledge of the explanatory variables. Sometimes predictor is used more restrictively to mean the explanatory variables that are used to make the prediction.

probabilistic forecast.

A forecast that specifies the future probability $\hat{p}(x)$ of one or more events x occurring. The set of events can be discrete (categorical) or continuous. **Deterministic forecasts** can be considered to be the special case of probability forecasts in which the forecast probabilities are always either zero or one - there is never any prediction uncertainty in the predictand. However, it is

perhaps more realistic to consider deterministic forecasts to be forecasts in which the prediction uncertainty in the predictand is not supplied as part of the forecast rather than as ones in which the prediction uncertainty is exactly equal to zero. Subjective probability forecasts can be constructed by eliciting expert advice.

probability of detection (POD).

Same as **hit rate (H)**.

proper/propriety.

A metaverification property for screening of suitable scores for probabilistic forecasts. A **strictly proper** score is one for which the best expected score is only obtained when the forecaster issues probability forecasts consistent with their beliefs (e.g. the forecasting model is correct). Proper scores discourage forecasters from *hedging* their forecasted probabilities towards probabilities that are likely to score more highly. The **Brier score** is **strictly proper**.

ranked probability score (RPS).

An extension of the Brier probability score to probabilistic categorical forecasts having more than two *ordinal* categories. By using cumulative probabilities, it takes into account the ordering of the categories.

refinement

Refinement is a statistical property of the forecasts that has multiple definitions in the verification literature. It can mean the marginal probability distribution of the forecasts, $p(\hat{x})$, as used in the phrase *calibration-refinement factorisation* (see **joint distribution**). However, often it is more specifically used to refer to the spread of the marginal probability distribution of the forecasts. In addition, refinement is also used synonymously to denote the **sharpness** of probability forecasts. However, in the Bayesian statistical literature refinement appears to be defined somewhat differently to sharpness (see DeGroot and Fienberg 1983).

reliability.

The same as **calibration**. It is related to Type 1 **conditional bias** $E_X(X \mid \hat{X}) - \hat{X}$ of the observations given the forecasts. Systems with zero conditional bias for all $\hat{X}$ are *perfectly reliable (well-calibrated)* and so have no need to be recalibrated (bias corrected) before use. Forecasting systems can be made more reliable by posterior recalibration; for example, the transformed forecast quantity $\hat{X}' = E_X(X \mid \hat{X})$ is perfectly reliable. Similar ideas also apply to probabilistic forecasts where predictors $\hat{X}$ are replaced by forecast probabilities $\hat{p}(x)$.

reliability diagram.

A diagram in which the conditional expectation of a predictand for given values of a continuous predictor is plotted against the value of the predictor. For deterministic forecasts of continuous variables, this is a plot of $f(\hat{x}) = E_X(X \mid \hat{X} = \hat{x})$ versus $\hat{x}$ i.e. the means of the observations stratified on cases with specific forecast values versus the forecast values. For probabilistic forecasts of binary events, this is a plot of $f(q) = E_X(X \mid \hat{p} = q)$ versus the forecast probability value q. Perfectly reliable forecasts have points that lie on the line $f(\hat{x}) = \hat{x}$ (deterministic forecasts of continuous variables) or $f(q) = q$ (probabilistic forecasts of binary variables). Given enough previous forecasts, recalibrated future forecasts can be obtained to good

approximation by using the reliability curve to non-linearly transform the forecasts: $\hat{x}' = f(\hat{x})$ or $\hat{p}' = f(\hat{p})$.

resistant measure.

A verification measure not unduly influenced by the presence of very large or small outlier values in the sample e.g. Mean Absolute Deviation.

resolution.

The sensitivity of the conditional probability $p(x \mid \hat{x})$ to different forecast values of $\hat{x}$. If a forecasting system leads to identical probability distributions of observed values for different forecast values, i.e. $p(x \mid \hat{x}_1) = p(x \mid \hat{x}_2)$, then the system has no resolution. Resolution is essential for a forecasting system to be able to discriminate observable future events. A single overall summary measure is provided by the variance of the conditional expectation $\mathrm{var}_{\hat{X}}[E_X(X \mid \hat{X})]$.

robust measure.

A verification measure that is not overly sensitive to the form of the probability distribution of the variables.

ROC.

A Relative (or Receiver) Operating Characteristic (ROC) is a signal detection curve for binary forecasts obtained by plotting a graph of the **hit rate** (y-axis) versus the **false alarm rate** (x-axis) over a range of different thresholds. For deterministic forecasts of a continuous variable, the threshold is a value of the continuous variable used to define the binary event. For probabilistic forecasts of a binary event, the threshold is a probability decision threshold that is used to convert the probabilistic binary forecasts into deterministic binary forecasts.

root mean square error (RMSE).

The square root of the **mean square error**.

sharpness.

An attribute of the marginal distribution of the forecasts that aims to quantify the ability of the forecasts to "stick their necks out". In other words, how much the forecasts deviate from the mean climatological value/category for deterministic forecasts, or from the climatological mean probabilities for probabilistic forecasts. Unvarying climatological forecasts take no great risks and so have zero sharpness; perfect forecasts are as sharp as the time-varying observations. For deterministic forecasts of discrete or continuous variables, sharpness is most simply estimated by the variance $\mathrm{var}(\hat{X})$ of the forecasts. For **perfectly calibrated** forecasts where $E_X(X \mid \hat{X}) = \hat{X}$, the sharpness $\mathrm{var}(\hat{X})$ becomes identical to the **resolution** $\mathrm{var}_{\hat{X}}[E_X(X \mid \hat{X})]$ of the forecasts. For probabilistic forecasts, although sharpness can also be defined by the variance $\mathrm{var}(\hat{p})$, it is often frequently defined in terms of the *information content (negative entropy)* $I = E(\hat{p} \log \hat{p})$ of the forecasts. High-risk forecasts in which $\hat{p}$ is either 0 or 1 have maximum information content and are said to be *perfectly sharp*. **Perfectly calibrated** perfectly sharp forecasts correctly predict all events. By interpreting deterministic forecasts as probabilistic forecasts with zero prediction uncertainty in the predictand, deterministic forecasts may be considered to be perfectly sharp probabilistic forecasts. However, it is perhaps more realistic to consider deterministic forecasts to be ones in which the prediction uncertainty in the predictand is not supplied as part of the forecast rather than ones in which the prediction uncertainty is

exactly equal to zero. Hence, a deterministic forecast can be considered to be a deterministic forecast with spread/sharpness $\mathrm{var}(\hat{X})$, yet at the same time can also be considered to be a probability forecast with perfect sharpness. The word **refinement** is also sometimes used to denote sharpness.

skill score.

Relative measure of the quality of the forecasting system compared to some (usually "low-skill") benchmark forecast. Commonly used reference forecasts include mean climatology, persistence (random walk forecast), or output from an earlier version of the forecasting system. There are as many skill scores as there are possible scores and they are usually based on the expression

$$SS = \frac{S - S_0}{S_1 - S_0} \times 100\%$$

where S is the forecast score, S_0 is the score for the benchmark forecast, and S_1 is the best possible score. The skill scores generally lie in the range 0 to 1 but can in practice be negative when using good benchmark forecasts (e.g. previous versions of the forecasting system). Compared to raw scores, skill scores have the advantage that they help take account of non-stationarities in the system to be forecast. For example, improved forecast scores often occur during periods when the atmosphere is in a more persistent state.

success ratio (SR).

A categorical binary score equal to the number of hits divided by the total number of events predicted $a/(a+b)$. Conditioned on the forecasts unlike the **hit rate**, which is conditioned on the observations.

sufficiency.

The concept of sufficiency was introduced into forecast evaluation by DeGroot and Fienberg (1983), and developed by Ehrendorfer and Murphy (1988) and Krzysztofowicz and Long (1991b) among others. When it can be demonstrated, sufficiency provides an unequivocal ordering on the quality of forecasts. When two forecasting systems, A and B say, are being compared, A's forecasts are said to be sufficient for B's if forecasts with the same skill as B's can be obtained from A's by a stochastic transformation. Applying a stochastic transformation to A's forecasts is equivalent to randomizing the forecasts, or passing them through a noisy channel (DeGroot and Fienberg 1983). Note that sufficiency is an important property in much of statistical inference, but the usage of the term is somewhat different in that context.

threat score (TS).

Same as **critical success index (CSI)**.

uncertainty.

The mean spread in the observations related to the width of the marginal probability distribution $p(x)$. Uncertainty is most simply measured by the variance, $\mathrm{var}(X)$, of the observations. Important aspect in the performance of a forecasting system, over which the forecaster has no control.