430

Recommendations on the verification of local weather forecasts

Pertti Nurmi[1]

Operations Department

[1] Finnish Meteorological Institute

December 2003

The Library
ECMWF
Shinfield Park
Reading, Berks RG2 9AX

library@ecmwf.int

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/library/ecpublications/

1. Introduction - Background

The ECMWF Technical Advisory Committee (TAC) noted at its 32nd session (2002) that the "Recommendations on the verification of local weather forecasts" annexed to the annual Report on Verification of ECMWF products in Member States and Co-operating States (hereafter referred to as MS), the so-called "Green Book", had been drafted some ten years ago. The TAC therefore requested that these recommendations be reviewed and revised in the light of current circumstances.

Recent progress in numerical weather prediction, as well as developments in forecast verification methods has been vigorous. The advent of probabilistic methods into operational numerical weather prediction has taken place during the last decade, and with the introduction of the Ensemble Prediction Systems (EPS) dramatically widened the use and applicability of NWP output in operational weather services within ECMWF MSs.

There are, and have been, various verification activities under the auspices of WMO like the newly founded Working Group on Verification (WGV) ([web 1]) within the World Weather Research Program (WWRP), or the more established verification group under the Working Group on Numerical Experimentation (WGNE) (Bougeault, 2003; [ref 1]). The emphasis of the latter is focused on verification techniques oriented toward model developers, while the role of the WGV is more directed to end users of high impact weather forecasts.

There is a host of recent important international conferences and workshops, either solely dedicated to verification issues, e.g.

- Workshop on Making Verification More Meaningful (Boulder, 2002; [ref 2], [web 2])
- WWRP/WMO Workshop on the Verification of Quantitative Precipitation Forecasts (Prague, 2001; [web 3])
- EUMETNET/SRNWP Mesoscale Verification Workshop (De Bilt, 2001; [ref 3])

or, with a strong verification context, e.g.

- International Conference on Quantitative Precipitation Forecasting (Reading, 2002; [ref 4])
- The biennial European Conference(s) on Applications of Meteorology (ECAM)

Two important textbooks with wide coverage on forecast verification methodologies need be highlighted, the earlier by Wilks (1995; [ref 5]) and the very recent by Jolliffe and Stephenson (2003; [ref 6]). A historical survey on verification methodology was compiled by Stanski et al. (1989; [ref 7]).

The Internet has dramatically established itself as the media and the means to communicate information. There are many websites with a wealth of verification content and their value is undeniable (e.g. [web 4, 5, 6]). However, one is easily lost in the web space where various different notations and formulae flourish depicting same methods and measures.

The past few years have seen efforts in harmonizing international verification practices. Strict rules to slavishly follow pre-defined verification measures and scores has proven to be a difficult and an undesirable task. Nevertheless, it is strongly advisable to adopt a general, coherent framework in forecast verification and to utilize common state-of-the-art methods. One example toward this objective was the WMO/CBS realized Standardised Verification System for Long-Range Forecasts ([web 7]). For purely model-based large-scale numerical forecasts standardisation is, however, fairly straightforward compared to harmonizing the verification of various local weather forecast products, originating at operational national weather offices,

where forecasting practices, parameters, lead times, forecast lengths, valid periods etc. are typically quite different.

Most of the above has taken place since the previous ECMWF "Green Book" verification recommendations were produced. A revision is therefore justified. It is the objective of these updated recommendations to take into account recent developments and guidelines in verification and also to cope with new model developments and forecast products originating thereof, without neglecting the common traditional methods.

The original reasoning and ideology behind the recommendations and the eventual "Green Book" contributions by the MSs have, however, not changes in the course of time. The previous reports and the existing "verification history" they contain serve as a valuable reference for future reports. The reports are meant as a forum to provide, on the one hand, valuable **exchange information between the MSs** to learn from each others' experiences and, on the other hand, to **produce valuable feedback to the Centre** on MS's verification activities and results of localized model behaviour, and even to distinguish possible model weaknesses. The latter function does not necessarily fall into the primary activities of the ECMWF itself where a more global verification approach is applied.

Chapter 2 of the recommendations provides some general guidelines, followed by an overview of the properties of various verification measures for continuous meteorological variables (Chapter 3), for binary and multi-category weather events (Chapter 4) and for probabilistic forecasts (Chapter 5). Forecast value and the end user decision making issues associated with forecast verification is covered briefly in Chapter 6, followed by a short Chapter 7 on other related issues concerning MSs verification activities. Proposals for means and measures to be followed up in MSs' annual contributions to the "Green Book" are highlighted and proposed at the end of each chapter.

The recommendations are outlined, having taken into account what has been reported by MSs in the "Green Books" of recent years, and, when appropriate, to be in harmony with the latest textbook on verification ([ref 6]), where an interested reader is referred to. It is the idea to keep the proposal at a fairly simple level to enable and encourage easy and straightforward applicability. In addition, MSs are warmly welcome to contribute whatever local verification studies they may think of being of general interest. At the end of the document, there are two lists of references, one to printed literature (quoted by [ref #] in the text) and, the other, for recommended websites existing at the time of writing (quoted by [web #]).

It is planned that these recommendations will eventually find their way under the ECMWF website (probably as a downloadable "pdf" document), where additions and possible corrections can be applied. The web version is meant as a helpful, living guidance when the preparation of national verification contributions is topical.

2. General guidelines

While the ECMWF boasts a comprehensive system to perform standard verifications of the upper air fields, the emphasis of the requested MS reporting is on the **verification of local forecasts of weather elements and (severe) weather events**. The origin of such forecasts may be the relevant parameters based on ECMWF direct model output (**DMO**). A natural second origin would be statistically or otherwise adapted, post-processed products (**PPP**) basing, e.g. on local perfect prog, MOS, or Kalman filtering schemes. The third forecast source would be the End Products (**EP**) delivered to the final end users. Although ECMWF is essentially aiming at medium-range (and longer) forecast ranges, it is appropriate and encouraged to produce **comparisons** of ECMWF DMO and derived PPP against corresponding output deriving from local numerical models like national Limited Area Models. Thus, an obvious comparison of a forecast production chain would comprise of:

DMO (model $_i$) vs. PPP (model $_i$) vs. EP,

where subscript $_i$ defines the model (ECMWF,…)

An analysis would then be obtained of the local post-processing scheme's ability to add value to direct model output and, additionally, whether local forecasters are able to outperform either guidance.

Since the ECMWF output is being disseminated in various horizontal grid resolutions and because MSs are possibly applying various of these (e.g. 0.5 vs. 1.5 degrees) in their applications and, further, because local models presumably also have various resolutions, it is requested to report on the **grid resolution** that has been used in the relevant verification statistics. Somewhat addressing this issue is the so-called "**double penalty**" problem, i.e. objective verification scores for local weather parameters may be better for a low resolution model than for a high resolution model. Although increased resolution typically provides more detailed small scale structures and stronger gradients in the forecasts, the consequent space and timing errors will easily be superfluous as compared to a lower resolution model. Especially if the scoring methods involve squared error measure (like the RMSE) the results may be quite misleading. One should try to elaborate this feature in the interpretation of the eventual verification statistics.

The verification process involves as one of its most central features the definition of the true state of the observed weather. Likewise forecasts, uncertainties and errors are evident in the observations. Traditionally, the observations originate from the synop observing network. It is, however, encouraged to adopt and experiment with new, more unconventional and more detailed observational data like those of meteorological **radars and satellites** as the observational "truth" in forecast verification.

With the increase in the resolution of numerical models it may be the case that model resolution exceeds that of the observations, leading to an inherent verification dilemma. The horizontal scale difference between observations and forecasts remains easily neglected. The density of the (traditional) observing network is highly variable. This raises the question of **point vs. area-averaged verification**. When the resolution of observations is higher than that of the model to be verified, one can **upscale** (e.g. Cherubini et al., 2001; [ref 8]) the observations to the model grid, rather than compute verification statistics against synop stations nearest to individual model gridpoints. This has proven to give more realistic and justified verification statistics. On the other hand, when the model resolution exceeds that of the observations, the **closest gridpoint** approach is often preferable. Care must be taken, however, close to coastlines or in variable terrain. Approaches to increase the availability and **representativeness of observational data** is in all cases of utmost importance.

The basic general framework of forecast verification addresses to the **joint distribution** of forecast vs. observation pairs and the methods to perform comparisons between them. A deterministic or a probabilistic (dichotomous or multivariate) distribution, **p** (**f**orecasts, **o**bservations), can be split into marginal distributions of forecasts, **p (f),** and observations, **p (o)**, and, further, the conditional distributions of forecasts given observations, **p (f|o)**, and observations given forecasts, **p (o|f)**. More of the subject can be found in an important paper by Murphy and Winkler (1987; [ref 9])

The **aggregation** of forecast vs. observation pairs into sufficiently large samples for evaluation is often required (for statistical significance) but, inversely, **stratification** of the results to be able to distinguish revealing details in the behaviour of the forecasts (or the models) is equally or even more important. There are various foundations for stratification:

- **time**; annual, biannual, seasonal, quarterly, monthly, time of day (diurnal cycle)
- **forecast range**; degradation of scores with lead time

- **values of the quantity or thresholds of the event**
- **spatial**; effects of land-sea contrast, altitude, snow-covered vs. bare terrain etc.

A comprehensive verification system will include a **reference** no-skill forecasting system against which to compare the forecasts. **Climatology, persistence** and **chance** are examples of references needed for the computation of the **skill score** and the **economic value**. Persistence typically provides a more competitive reference forecasts than climate up to c. two days forecast range. Both should be quite easily derived within national weather services, so utilization of **both references** is proposed. Likewise, the verification of probabilistic forecasts requires knowledge of the **climatological distributions** or **cumulative probability distributions** (cdf) of the relevant events. From the model point of view the Centre has a relatively sound knowledge of **model climate**. However, the MSs, having access to their own observation databases, are in a more proper position to define **local observation-based climatological distributions** to produce reference verification data both in the measurement and in the probability space.

Verification statistics should be accompanied by **statistical significance testing**, especially in the cases of severe/extreme weather events. The relative frequency of extreme weather is, by definition, very low and, consequently, sample sizes small. Wrong conclusions are therefore easily being made. Extreme event forecasting should be supported by **probabilistic guidance** like the ECMWF **Extreme Forecast Index (EFI)**.

The MSs are strongly encouraged to develop **operational, online, real-time verification software** with a modular structure for easy updates and modifications. An added facility to produce **periodical verification reports** covering the most common verification measures is likewise supported. Such software already exists in a number (~10) of MSs according to their "Green Book" reporting. Operational verification packages enable a fairly straightforward reproduction of verification statistics to serve the additional purpose of contributing to the "Green Book" on a regular, coherent basis. It is requested to continue keeping ECMWF (and other MSs) informed whether (i) operational verification schemes (either intra- or internet) exist and/or, (ii) periodical verification reports are being produced.

To summarize, it is proposed to:

- verify local forecasts of weather elements and severe weather events
- compare DMO vs. PPP vs. EP
- consider model grid resolution(s) being used
- evaluate the representativeness of observational data
- distinguish outliers in data
- derive local climatological distributions, including cumulative probability distributions
- apply radar and/or satellite observations in addition to conventional observational data
- consider point vs. area verification, taking into account upscaling of observations and the closest gridpoint approach
- utilize several no-skill reference forecasts to compute verification scores
- perform aggregation **and** stratification of results
- perform statistical significance and hypothesis testing
- compute and analyse the economic value of forecasts
- develop operational verification systems and report on their features

3. Continuous variables

The verification of continuous variables typically provides statistics on how much the forecast values differ from the observations and, thereafter, computation of relative measures against some reference forecasting systems. The most common continuous local weather parameters to verify are:

- **Temperature**: fixed time (e.g. noon, midnight), Tmin, Tmax, time-averaged (e.g. five-day)
- **Wind speed and direction**: fixed time, time-averaged
- **Accumulated precipitation**: time-integrated (e.g. 6, 12, 24 hours)
- **Cloudiness**: fixed time, time-averaged; typically categorized

Their behaviour can, however, be quite different: when the temperature may behave quite smoothly and follow a Gaussian distribution, the wind speed is often very sporadic, the precipitation intermittent, and the cloudiness following a U-shaped distribution.

The best first way to approach verification of continuous predictands is to produce **scatter plots** of forecasts vs. observations. Rather than being a verification measure, scatterplot is a means to explore the data and can thus provide a visual insight to the correspondence between forecast and observed distributions. An excellent feature is the possibility to distinguish at a glance potential **outliers** either in the forecast or in the observation dataset. Accurate forecasts would have the points lined on a 45 degree diagonal in a square scatterplot box. Additional useful ways to produce scatterplots are in the form of:

- **observation vs. [forecast - observation]**
- **forecast vs. [forecast - observation]**

i.e. either the observation or the forecast plotted against their difference. Such plotting provides a visually descriptive method to see how forecast errors behave with respect to observed or forecast distributions revealing potential clustering or curvature in their relationships.

In a similar manner as the scatterplot, a **time-series plot** of forecasts vs. observations (or forecast error) quite easily uncovers potential outliers in either forecast or observation datasets. **Trends** and **time-dependent relationships** are easily discernible. Neither scatterplots nor time series plots will provide any concrete measures of accuracy.

The next proposed step is always to compute the simple average difference between the forecast and the observation, the **systematic** or the **Mean Error (bias)**:

$$\mathbf{ME = (1/n) \ \Sigma (f_i - o_i)}$$

The bias is the simplest and most familiar of scores and can provide very useful information on the local behaviour of a given weather parameter (e.g. maximum temperature close to the coastline or minimum temperature over snow-covered ground). The ME range is from minus infinity to infinity, and a perfect score is = 0. However, it is possible to reach a perfect score for a dataset with large errors, if there are compensating errors of a reverse sign. The ME is not an accuracy measure as it does not provide information of the magnitude of forecast errors.

A simple measure to compensate for the potential positive and negative errors of the ME is to next compute the **Mean Absolute Error**:

$$\mathbf{MAE = (1/n) \ \Sigma \ | f_i - o_i |}$$

The MAE range is from zero to infinity and, as with the ME, a perfect score equals = 0. The MAE measures the average magnitude of forecast errors in a given dataset and therefore is a scalar measure of forecast accuracy. **It is advisable to always view the ME and the MAE simultaneously**.

Another common accuracy measure is the **Mean Squared Error**:

$$\mathbf{MSE = (1/n) \ \Sigma (f_i - o_i)^2}$$

or its square root, the **RMSE**, which would have the same unit as the forecast parameter. As with the MAE, their range is from zero to infinity with a perfect score of = 0. MSE is the squared difference between forecasts and observations. Due to the second power, the MSE and RMSE are much more sensitive to large forecast errors than the MAE. This may be especially harmful in the presence of potential outliers in the datasets and, consequently, at least with small or limited datasets the use of the MAE is preferred. The fear for the high penalty of large forecast errors will easily lead a forecaster to a conservative forecasting practice. MAE is also more practical from the duty forecasters' intuition as it shows the errors in the same unit and scale as the parameter itself.

A recommended (at least for experimentation) measure which, however, is not yet in wide use is the **Linear Error in Probability Space**:

$$\mathbf{LEPS = (1/n) \ \Sigma \mid CDF_o (f_i) - CDF_o (o_i) \mid ,}$$

where CDF_o is the Cumulative probability Density Function of the observations, determined from a relevant climatology. (Note: LEPS should not be confused with another, completely different LEPS notation, the Limited-area Ensemble Prediction System!) LEPS is the MAE in probability, rather than measurement space, and is defined as the mean absolute difference between the cumulative frequency of the forecast and the cumulative frequency of the observation. Its range is from zero to unity, with a perfect score equalling = 0. LEPS does not depend on the scale of the variable to be verified and takes the variability of the parameter into account. It can be used to evaluate forecasts between different locations. LEPS computation may require some elaboration of the local observation datasets because of the need for appropriate climatological cumulative distributions at each forecast point. Thereafter its derivation is straightforward. Nevertheless, this is much more natural to be done **locally at MSs** than by the ECMWF. An attractive feature of the LEPS is that it encourages forecasting in the extreme tails of the climate distributions, when justified, by penalizing less than for a similar size error in a more probable region of the climatological distribution.

The original form of LEPS is reported to "exhibit certain pathological behaviour at its extremes" ([ref 6, p. 92). Therefore certain correction and normalization terms have been introduced, leading to:

$$\mathbf{LEPS_{rev} = 3* (1 - \mid F_f - F_o \mid + F_f^{\,2} - F_f + F_o^{\,2} - F_o) - 1} \ , \quad \text{where}$$

F_f and F_o are the CDF_os of the forecasts and observations, respectively.

Relative accuracy measures that provide estimates of the (percentage) improvement of the forecasting system over a reference system can be defined in the form of a general **skill score**:

$$\mathbf{SS = (A - A_{ref}) / (A_{perf} - A_{ref})} \ ,$$

where A = the applied measure of accuracy, A_{perf} = the value of the accuracy measure which would result from perfect forecasts, and A_{ref} = the accuracy value of reference forecasts, typically climatology or persistence (both should be used). For negatively oriented accuracy measures (i.e. smaller values of A are better, like MAE, LEPS, and MSE) the skill score becomes:

$$\mathbf{SS = 1 - A / A_{ref}}$$

It is encouraged to compute the skill of EP vs. PPP vs. DMO. Consequently, it is proposed to apply:

MAE_SS = 1 - MAE / MAE_{ref}

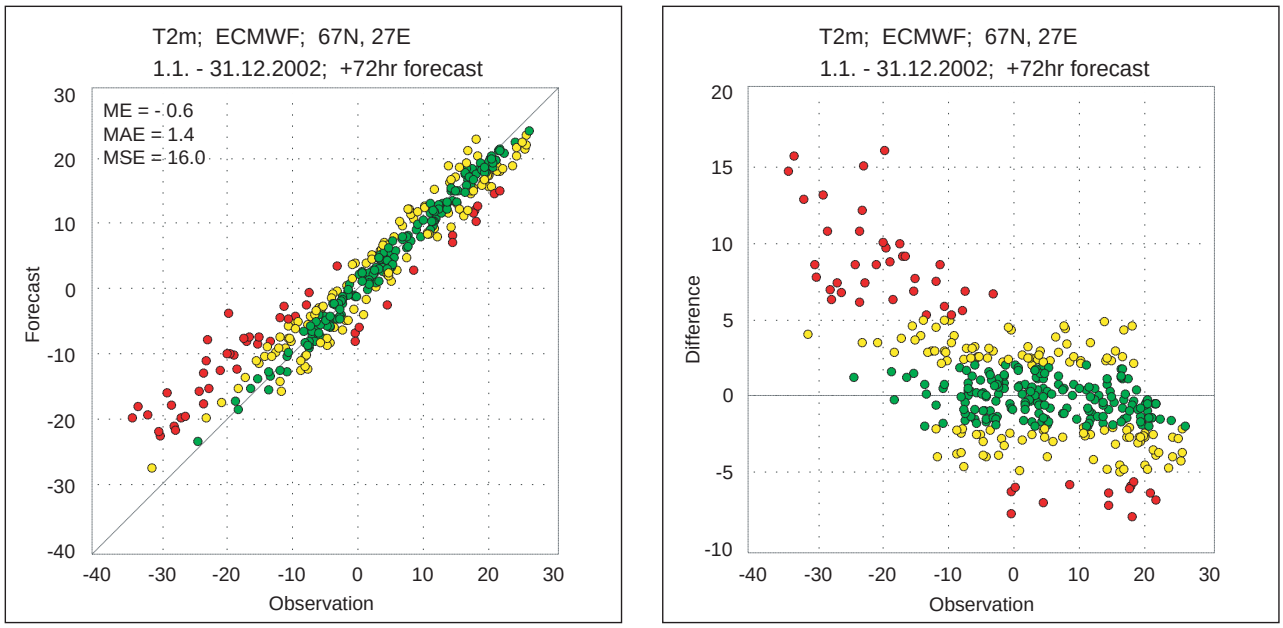
LEPS_SS = 1 - LEPS / $LEPS_{ref}$

MSE_SS = 1 - MSE / MSE_{ref}

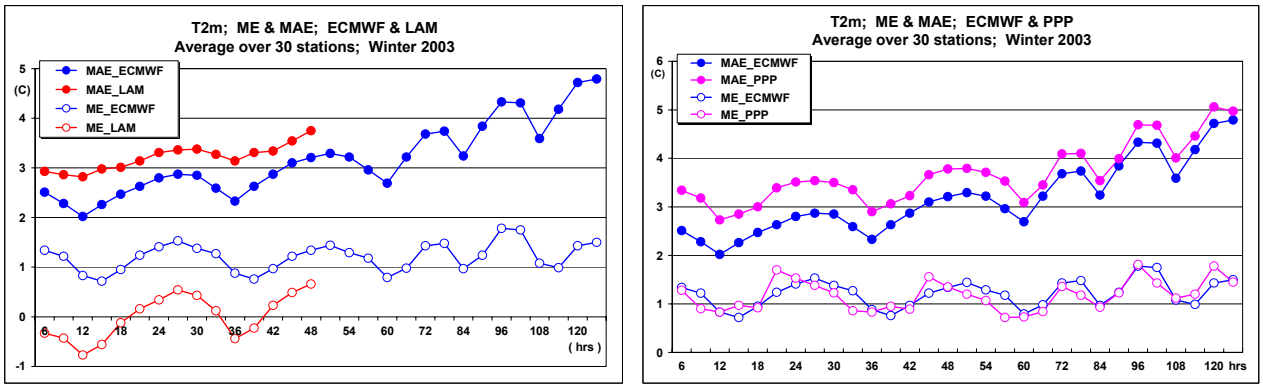
The range of skill scores is minus infinity to unity (for a perfect forecast system), with a value = 0 indicating no skill over the reference forecasts. Skill scores can be unstable for small sample sizes, especially if MSE_SS were used.

To summarize (including the general guidelines), and indicating minimum and optimum requirements, it is proposed to:

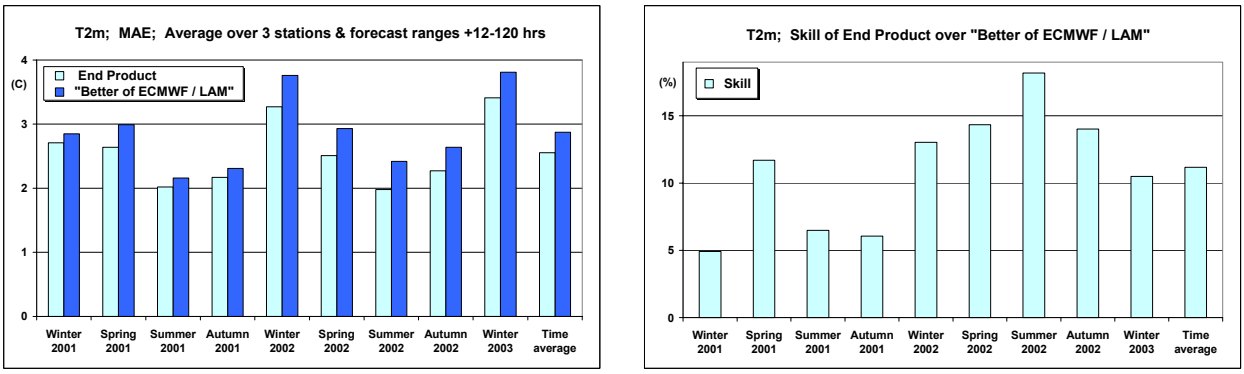
- verify a comprehensive set of continuous local weather variables
- minimum proposal: produce scatterplots and time-series plots, including forecasts and/or observations against their difference
- minimum proposal: compute ME, MAE, MAE_SS
- optimum proposal: compute LEPS (and $LEPS_{rev}$), LEPS_SS, MSE, MSE_SS



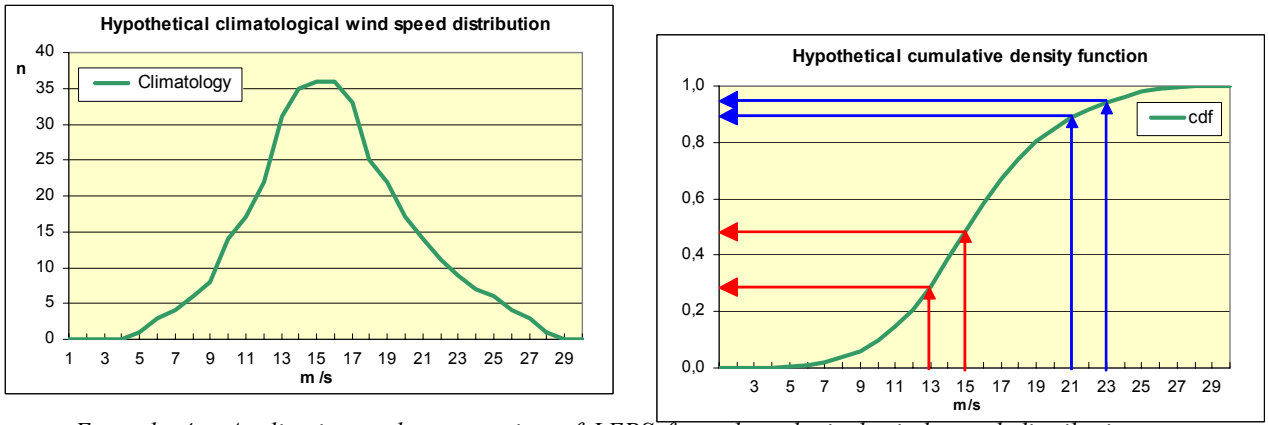
Example 1 - Scatterplot of one year of ECMWF three-day T2m forecasts (left) and forecast errors (right) versus observations at a single location. Red, yellow and green dots separate the errors in three categories. Some basic statistics like ME, MAE and MSE are also shown. The plots reveal the dependence of model behaviour with respect to temperature range, i.e. over- (under) forecasting in the cold (warm) tails of the distribution.



Example 2 - Temperature bias and MAE comparison between ECMWF and a Limited Area Model (LAM) (left), and an experimental post-processing scheme (PPP) (right), aggregated over 30 stations and one winter season. In spite of the ECMWF warm bias and diurnal cycle, it has a slightly lower MAE level than the LAM (left). The applied experimental "perfect prog" scheme does not manage to dispose of the model bias and exhibits larger absolute errors than the originating model – this example clearly demonstrates the importance of thorough verification prior to implementing a potential post-processing scheme into operational use.



Example 3 - Mean Absolute Errors of End Product and DMO temperature forecasts (left), and Skill of the End Products over model output (right). The better of either ECMWF or local LAM is chosen up to the +48 hour forecast range (hindcast), thereafter ECMWF is used. The figure is an example of both aggregation (3 stations, several forecast ranges, two models, time-average) and stratification (seasons).



Example 4 - Application and computation of LEPS for a hypothetical wind speed distribution at an assumed location, where the climatological frequency distribution (left) is transformed to a cumulative probability distribution (right). A 2 m/s forecast error around the median, in the example 15 m/s vs. 13 m/s (red arrows), would yield a LEPS value of c. 0.2 in the probability space (| 0.5 – 0.3 |, red arrows). However, an equal error in the measurement space close to the tail of the distribution, 23 m/s vs. 21 m/s (blue arrows), would result a LEPS value of c. 0.05 (| 0.95 – 0.9 |, blue arrows). Hence forecast errors of rare events are much less penalized using LEPS.

4. Categorical events

4.1 Binary (dichotomous; yes/no) forecasts

Categorical statistics are needed to evaluate binary, yes/no, forecasts of the type of statements that an event will or will not happen. Typical binary forecasts are warnings against adverse weather like:

- **Rain** (vs. no rain); with various rainfall thresholds
- **Snowfall**; with various thresholds
- **Strong winds** (vs. no strong wind); with various wind force thresholds
- **Night frost** (vs. no frost)
- **Fog** (vs. no fog)

The first step to verify binary forecasts is to compile a **2*2 contingency table** showing the frequency of "yes" and "no" forecasts and corresponding observations:

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	Hit	False alarm	Fc Yes
No	Miss	Correct rejection	Fc No
Marginal total	Obs Yes	Obs No	Sum total

=>

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d = n

There are two cases when the forecast is correct, either a "hit" or a "correct rejection" (or "correct no forecast") and two cases when the forecast is incorrect, either a "false alarm" or a "miss". The so-called marginal distributions of the forecasts and observations are the totals that are provided in the right columns and lower rows of the contingency tables, respectively. A perfect forecast system would have only hits and correct rejections, with the other cells being = 0. Occasionally one sees the tables transposed, i.e. forecast and observed cell counts reversed. The distribution above is clearly the more popular one in literature and should be utilized for harmony.

The seemingly simple definition of a binary event, and the subsequent 2*2 contingency table, hides quite astonishing complexity. There are a number of measures to tackle this complex issue and they are defined here highlighting some of their properties. Most, if not all, have a long historical background but they are still used very commonly. One should remember that **in no case is it sufficient to apply only just one single verification measure**.

The **Bias** of binary forecasts compares the frequency of forecasts (Fc Yes) to the frequency of actual occurences (Obs Yes) and is represented by the ratio:

$$\mathbf{B = (a + b) / (a + c)} \qquad [\sim \text{Fc Yes / Obs Yes}]$$

Range of B is zero to infinity, an unbiased score = 1. With B > 1 (< 1), the forecast system exhibits over-forecasting (under-forecasting) of the event. B is also known as **Frequency Bias Index (FBI).** As in the case of continuous variables, bias is not an accuracy measure.

The most simple and intuitive performance measure that provides information on the accuracy of a categorical forecast system is **Proportion Correct**:

$$\mathbf{PC = (a + d) / n} \qquad [\sim (\text{Hits} + \text{Correct rejections}) / \text{Sum total}]$$

Range of PC is zero to one, a perfect score = 1. PC is usually very misleading because it rewards correct "yes" and "no" forecasts equally and is strongly influenced by the more common category. This is typically the "no event" case, i.e. not the extreme event of interest.

The measure that examines by default the (extreme) event by measuring the proportion of observed events that were correctly forecast is **Probability Of Detection**:

POD = a / (a + c) [~ Hits / Obs Yes]

Range of POD is zero to one, a perfect score = 1. It is also called the **Hit Rate (H)** which should not be confused with PC. The complement of H (or POD) is the Miss Rate (i.e. 1 - H or c/(a+c)) which gives the relative number of missed events. POD is sensitive to hits but takes no account of false alarms. It can be artificially improved by producing excessive "yes" forecasts to increase the number of hits (with a consequence of numerous false alarms). While maximising the number of hits and minimizing the number of false alarms is desirable, it is required that POD be examined together with **False Alarm Ratio**:

FAR = b / (a + b) [~ False alarms / Fc Yes]

Range of FAR is one to zero, a perfect score = 0, i.e. FAR has a negative orientation. FAR is also very sensitive to the climatological frequency of the event. Contrary to POD, FAR is sensitive to false alarms but takes no account of misses. Likewise POD, it can be artificially improved, but now by producing excessive "no" forecasts, i.e. to reduce the number of false alarms. Because the increase of POD is achieved by increasing FAR and decrease of FAR by decreasing POD, **POD and FAR must be examined together**.

While FAR above is a measure of false alarms given the forecasts (Fc Yes), another score applying the cell counts of false alarms, **False Alarm Rate (note the difference in notation!)** is a measure of false alarms given the event did not occur (Obs No) (also known as **Probability Of False Detection, POFD**), and is defined as:

F = b / (b + d) [~ False alarms / Obs No]

Range of F is again one to zero, a perfect score = 0, i.e. like FAR exhibiting negative orientation. F is generally associated with the evaluation of probabilistic forecasts by combining it with POD (or H) into the so-called **Relative Operating Characteristic** diagram or curve (**ROC**, see Chapter 5). However, it is possible to apply the ROC in a categorical binary case so that one can compare directly and consistently a categorical forecast (point value) with a probability forecast (curve).

If a verification system covers computation of POD and F, a popular skill score with various "inventors" in the history is automatically generated: **Hanssen-Kuipers Skill Score (KSS)**, or **True Skill Statistics (TSS)**, or **Peirce Skill Score (PSS)**, is defined (in its simplest form) as:

KSS = POD - F (= H - F) [~ (Hits / Obs Yes) - (False alarms / Obs No)]

Range of KSS is minus one to one, a perfect score = 1, no skill forecast = 0 (i.e. POD = F). Ideally, KSS measures the ability of the forecast system to separate the "yes" cases (POD) from the "no" cases (F). For rare events, the frequency of correct rejections cell (d) is typically very high in the contingency table compared to the other cells, leading to a very low False Alarm Rate and, consequently, KSS is close to POD.

A widely used performance measure of rare events, is **Threat Score (TS),** or **Critical Success Index (CSI):**

TS = a / (a + b + c) [~ Hits / (Hits + False alarms + Misses)]

Range of TS is zero to one, a perfect score = 1, no skill forecast = 0. TS is sensitive to hits and takes into account both false alarms and misses and can be seen as a measure for the event being forecast after

removing correct (simple) "no" forecasts from consideration. TS is sensitive to the climatological frequency of events (producing poorer scores for rarer events), since some hits can occur due to random chance. To overcome this effect, a kindred score, **Equitable Threat Score** (also known as **Gilbert's Skill Score, GSS**) adjusts for the number of hits associated with random chance, and is defined as:

ETS = (a - ar) / (a + b + c - ar) [~ (Hits - Hits random) / (Hits + False alarms + Misses - Hits random)]

where **ar = (a + b) (a + c) / n** [~ (Fc Yes) * (Obs Yes) / Sum total]

is the number of hits for random forecasts.

Range of ETS is -1/3 to one, a perfect score = 1, no skill forecast = 0.

One of the most commonly used skill scores for summarizing the 2*2 contingency table is **Heidke Skill Score**. It's reference accuracy measure is Proportion Correct (PC), adjusted to eliminate forecasts which would be correct due to random chance. Using the cell counts it can be written in the form:

HSS = 2 (ad - bc) / { (a + c)(c + d) + (a + b)(b + d) }

Range of HSS is minus infinity to one, a perfect score = 1, no skill forecast = 0.

Odds Ratio measures the forecasting system's probability (odds) to score a hit (POD or H) as compared to the probability of making a false alarm (POFD or F):

OR = { H / (1 - H) } / { F / (1 - F) }, which using the cell counts becomes:

OR = ad / bc [~ (Hits * Correct rejections) / (False alarms * Misses)]

Range of OR is zero to infinity, a perfect score yields infinity, no skill system = 1, i.e. the ratio is greater than one when POD exceeds the False Alarm Rate. Odds Ratio is independent of potential biases between observations and forecasts because it does not depend on marginal totals of the contingency table. It can be transformed into a skill score, ranging from -1 to +1:

ORSS = (OR – 1) / (OR + 1), and using the cell counts:

ORSS = (ad - bc) / (ad + bc)

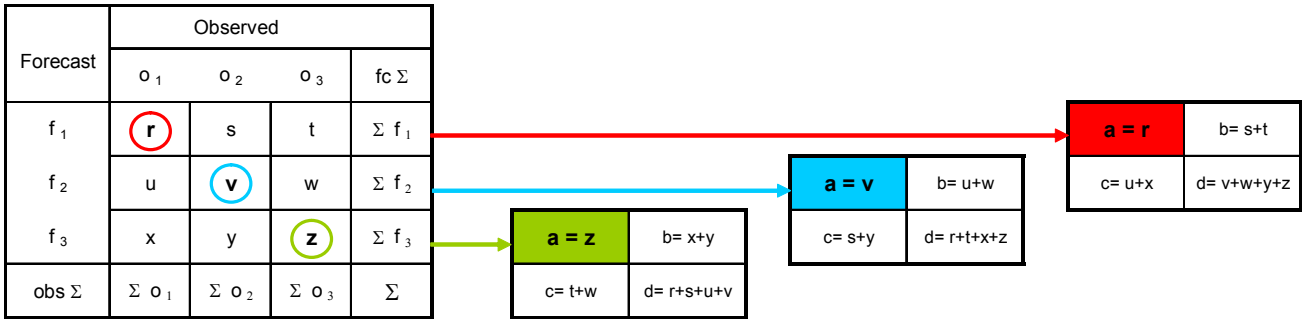
ORSS has practically never been used in meteorological forecast verification but is supposed to possess several attractive properties (Stephenson, 2000; [ref 10]). Because of this and simplicity of computation, it's use is proposed at least for experimentation.

4.2 Multi-category forecasts

Categorical events are naturally not limited to binary forecasts of two categories and the associated 2*2 contingency tables. The general distributions approach in forecast verification studies the relationship among the elements in multi-category contingency tables. One can consider local weather variables in several mutually exhaustive categories, e.g. cloudiness or accumulated rainfall in k categories (where k>2), or rain type classified into rain/snow/freezing rain types (k=3), and likewise for wind warnings categorized into strong gale/gale/no gale (k=3), etc.

It is advisable to initiate verification again by constructing a contingency table where the frequencies of forecasts and observations are collected in relevant cells as illustrated in the attached table for a 3*3 category case (left-hand box) (adapted from [ref 5]). A perfect forecast system would (again) have all the entries along the diagonal (r, v, z, in the example), all other values being = 0. Only the Proportion Correct (PC) can directly be generalized to situations with more than two categories. The other verification measures of

Chapter 4.1 are valid only with the binary yes/no forecast situation. To be able to apply these measures, one must convert the k>2 contingency table into a series of 2*2 tables. Each of these is constructed by considering the "forecast event" distinct from the complementary "non-forecast event", which is composed as the union of the remaining k-1 events (right-hand sub-boxes of the table, where the same cell notation is used as in the previous table). The off-diagonal cells provide information about the nature of the forecast errors. For example, biases (B) reveal if some categories are under- or over-predicted, while PODs quantify the success of detecting the distinct categorical events.



The KSS and HSS skill scores can be generalized to multi-category cases:

$$\mathbf{KSS = \{ \Sigma\, p(f_i, o_i) - \Sigma\, p(f_i)\, p(o_i) \} \;/\; \{ 1 - \Sigma\, (p(f_i))^2 \},}$$

$$\mathbf{HSS = \{ \Sigma\, p(f_i, o_i) - \Sigma\, p(f_i)\, p(o_i) \} \;/\; \{ 1 - \Sigma\, p(f_i)\, p(o_i) \},}$$

where the subscript i denotes the dimension of the table, $p(f_i, o_i)$ represents the joint distribution of forecasts and observations (i.e. the diagonal sum count divided by the total sample size, the PC), and $p(f_i)$ and $p(o_i)$ are the marginal probability distributions of the forecasts and observations (i.e. row and column sums divided by the sum total), respectively. Both KSS and HSS are measures of potential improvement in the number of correct forecasts over random forecasts. The estimation of randomness (denominator) is the only difference between these two scores. For a 2*2 situation the equations reduce to the corresponding formulae shown in the previous chapter.

Rain forecast	Rain observed		
	Yes	No	fc Σ
Yes	52	45	97
No	22	227	249
obs Σ	74	272	346

~>

B	= 1.31	TS	= 0.44
PC	= 0.81	ETS	= 0.32
POD	= 0.70	KSS	= 0.53
FAR	= 0.46	HSS	= 0.48
F	= 0.17	OR	= 11.92
		ORSS	= 0.85

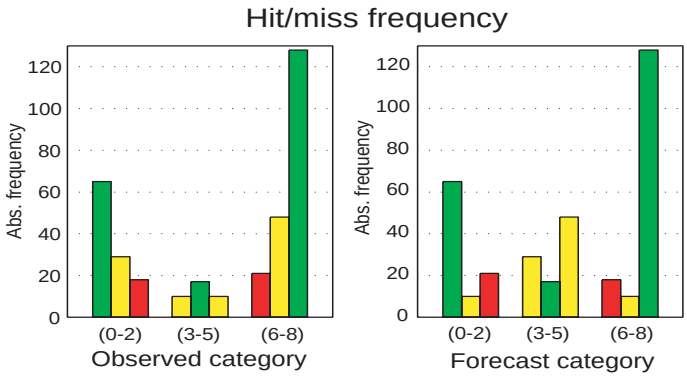
Example 5 - Contingency table of one year (with 19 missing cases) of categorical rain vs. no rain forecasts (left), and resulting statistics (right). Rainfall is a relatively rare event at this particular location, occurring in only c. 20 % (74/346) of the cases. Due to this, PC is quite high at 0.81. The relatively high rain detection rate (0.70) is "balanced" by high number of false alarms (0.46), with almost every other rain forecast having been superfluous. This is also seen as biased over-forecasting of the event (B=1.31). Due to the scarcity of the event the false alarm rate is quite low (0.17) – if used alone this measure would give a very misleading picture of forecast quality. The Odds Ratio shows that it was 12 times more probable to make a correct (rain or no rain) forecast than an incorrect one. The resulting skill score (0.85) is much higher than the other skill scores which is to be noted - this is a typical feature of the ORSS due to its definition.

Clouds forecast	Clouds observed			
	0 - 2	3 - 5	6 - 8	fc Σ
0 - 2	65	10	21	96
3 - 5	29	17	48	94
6 - 8	18	10	128	156
obs Σ	112	37	197	346

~>

No clouds (0-2)	Partly cloudy (3-5)	Cloudy (6-8)
B = 0.86 POD = 0.58 FAR = 0.32 F = 0.13 TS = 0.45	B = 2.54 POD = 0.46 FAR = 0.82 F = 0.25 TS = 0.15	B = 0.79 POD = 0.65 FAR = 0.18 F = 0.19 TS = 0.57

Overall: PC = 0.61 KSS = 0.41 HSS = 0.37



Example 6 - Multi-category contingency table of one year (with 19 missing cases) of cloudiness forecasts (left), and resulting statistics (right). Results are shown exclusively for forecasts of each cloud category, together with the overall PC, KSS and HSS scores. The most marked feature is the very strong over-forecasting of the "partly cloudy" category leading to numerous false alarms (B=2.5, FAR=0.8), and, despite this, the poor detection (POD=0.46). The forecasts cannot reflect the observed U shaped distribution of cloudiness at all. Regardless of this inferiority both overall skill scores are relatively high (c. 0.4), following the fact that most of the cases (90 %) fall either in the "no cloud" or "cloudy" category - neither of these scores takes into account the relative sample probabilities, but weight all correct forecasts similarly.
The lower part of the example shows the same data transformed into hit/miss bar charts, either given the observations (left), or given the forecasts (right). The green, yellow and red bars denote correct and one and two category errors, respectively. The U-shape in observations is clearly visible (left), whereas there is no hint of such in the forecast distribution (right).

To summarize (including the general guidelines), and indicating minimum and optimum requirements, it is proposed to:

- verify a comprehensive set of categorical events by compiling relevant contingency tables, including multi-category events, and focusing on adverse and/or extreme local weather
- minimum proposal: compute B, PC, POD, FAR, F, KSS, TS, ETS, HSS
- optimum proposal: compute OR, ORSS, ROC

5. Probability forecasts

All forecasting involves some level of uncertainty. However, deterministic forecasts and their verification in Chapters 3 and 4 do not address the inherent uncertainty of the weather parameter or event under consideration. Probabilistic forecasts, given probabilities of the expected event with values between 0 % and 100 % (or 0 and 1) much better take into account the underlying joint distribution between forecasts and observations. One should remember that a conversion of probability forecasts to categorical events is

possible and simple by just defining the "on/off" probability threshold. However, reverse is not straightforward. Verification of probability forecasts is, on the other hand, somewhat more laborious, not only because large datasets are required to obtain any significant information.

Probability forecasts can be produced with different methods just like categorical forecasts. We may have subjective probability forecasts to end users issued by forecasters (EP prob), or statistically post-processed probability forecasts (PPP prob), or forecasts generated from a set of deterministic numerical forecast like the ECMWF Ensemble Prediction System (EPS). Therefore, by using a similar notation as earlier in Chapter 2, it is possible and desirable to provide comparisons of the form:

EPS vs. PPP $_{prob}$ vs. EP $_{prob}$

A common first look at the behaviour of a probabilistic forecast system is to construct a **reliability diagram** (see Example 7, left). It represents an informative graphical plot of the observed relative frequency of an event as a function of it's forecast probability in definite probability categories (e.g. in 10% intervals). The resulting reliability curve is thus an indication of the agreement between mean forecast probability and mean observed frequency. Perfect reliability is reached when all forecast probabilities and corresponding observed relative frequencies are the same, aligned along the diagonal 45 degree line. The reliability diagram should include a summary distribution of the frequency of the use of each definite forecast probability category, which will depict the **sharpness** of the system. It indicates the capability of the system to forecast extreme values, or values close to 0 or 1. As with probability forecasts in general, the reliability diagram requires a large number of observation-forecast pairs to yield a meaningful diagram. A more comprehensive form of the reliability diagram is the so-called **attributes diagram** (see, [web 8]).

The most common measure of the quality of probability forecasts is the **Brier Score (BS)**. It measures the mean squared difference between forecasts and observations in probability space and is the equivalent of MSE of categorical forecasts. Likewise, it is negatively oriented, with perfect forecasts having BS = 0.

$$\mathbf{BS = (1/n) \ \Sigma (p_i - o_i)^2 ,}$$

where index i denotes the numbering of observation-forecast pairs, pi are the forecast probabilities of the given event and oi the corresponding observed values, having integer values 1 or 0, if the event occurred or did not, respectively. Analogous to earlier definitions, it is customary to generate a skill score, where a reference forecast system is required:

$$\mathbf{BS_{ref} = (1/n) \ \Sigma (ref_i - o_i) 2} ,$$

where ref_i is usually the relevant climatological relative frequency of the event.

The resulting **Brier Skill** Score is:

$$\mathbf{BSS = 1 - BS / BS_{ref}} .$$

The Brier Score can be algebraically decomposed into three quantities known as **reliability, resolution and uncertainty**. They are not elaborated here but, rather, reference is made to the User Guide to ECMWF Forecast Products ([ref 11], [web 9]) with illustrative examples.

A vector generalization of the Brier (Skill) Score to multi-event or multi-category situations is defined by the **Ranked Probability Score (RPS)** and the respective skill score. It measures the sums of squared differences in cumulative probability space for a multi-event probability forecast. It penalizes forecasts more severely when their probabilities are further from the actual observed distributions.

$$\mathbf{RPS = (1/(k-1)) \ \Sigma \{ (\Sigma p_i) - (\Sigma o_i) \}^2} ,$$

where k is the number of probability categories. Consequently:

$$\mathbf{RPSS = 1 - RPS / RPS_{ref}}$$

Both BSS and RPSS are very sensitive to dataset size.

Signal Detection Theory (SDT) has brought to meteorology a method to assess the performance of a forecasting system that distinguishes between the discrimination capability and the decision threshold of the system, namely the **Relative Operating Characteristic (ROC)**. This has attained wider and wider popularity in meteorological forecast verification during recent years. The **ROC curve** is a graphical representation in a square box of the Hit rate (H) (y-axis) against the False Alarm Rate (F) (x-axis) for different potential decision thresholds (see Example 7, right). H, rather than POD notation is used here to be consistent with the recent textbook in verification ([ref 6]). Graphically, ROC curve is plotted from a set of probability forecasts by stepping (or sliding) a decision threshold (e.g. with 10% probability intervals) through the forecasts, each probability decision threshold generating a 2*2 contingency table. Hence the probability forecast is transformed into a set of categorical "yes/no" forecasts. A set of value pairs of H and F is then obtained, forming the curve (For an explicit demonstration, see [ref 7, Chapter 4.1]). It is desirable that H be high and F be low. On the graph, the closer the point is to the upper left-hand corner, the better the forecast. Since a perfect forecast system would have only correct forecasts with no false alarms, regardless of the threshold chosen, a perfect system is represented by a ROC "curve" that rises from (0,0) (H=F=0) along the y-axis to (0,1) (upper left-hand corner; H=1, F=0) and then straight to (1,1) (H=F=1).

An attractive, relative index and widely used summary measure based on the diagram is the **ROC area ($\mathbf{ROC_A}$)**, the area remaining under the curve, and an area-based skill score (**ROC_SS**) derived from it. In a perfect forecast system ROCA would be =1. It decreases from one as the curve moves downward from the ideal top-left corner of the box. A useless, zero-skill, forecast system is represented as a straight line along the diagonal, when H=F and the area is = 0.5. Such a system cannot discriminate between occurences and non-occurences of the event. The ROCA based skill score can simply be defined as:

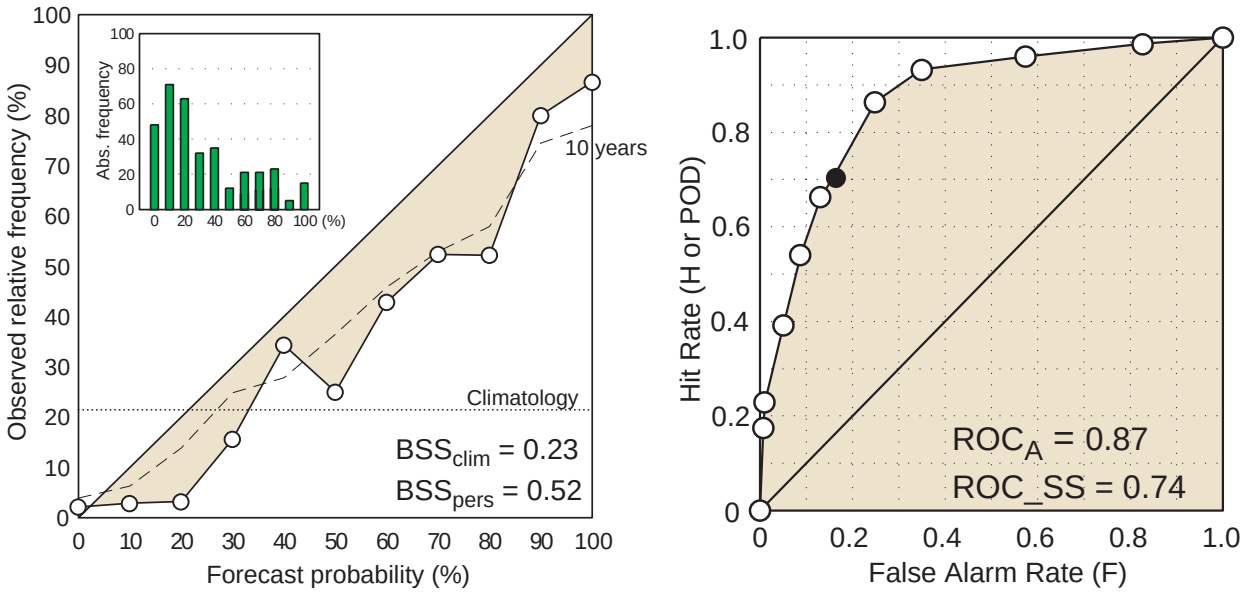
$$\mathbf{ROC\_SS = 2 * ROC_A - 1}$$

Below the diagonal ROC_SS has negative values, reaching a minimum of - 1, when ROCA equals = 0. It can be shown that for a deterministic forecast, ROC_SS translates into H - F, i.e. KSS.

As mentioned earlier in Chapter 4.1, ROC can be adapted for a categorical binary event. In that special case there is only one single decision threshold and, instead of a curve, only a single point results. An advantage of measures such as ROC, ROCA and ROC_SS is that they are directly related to a decision-theoretic approach and can thus be related to the economic value of probability forecasts for end users, and possibly allowing for the assessment of the costs of false alarms (see, Chapter 6).

To summarize (including the general guidelines), and indicating minimum and optimum requirements, it is proposed to:

- verify a comprehensive set of probability forecasts focusing on adverse and/or extreme local weather
- minimum proposal: produce reliability diagrams, including sharpness distribution
- minimum proposal: compute BS, BSS
- optimum proposal: produce attributes diagrams and ROC diagrams
- optimum proposal: decompose BS, compute RPS, RPSS, ROCA , ROC_SS



Example 7 - Reliability (left) and ROC (right) diagrams of one year of PoP (Probability of Precipitation) forecasts. The data are the same as in Example 5, where the PoPs were transformed into categorical yes/no forecasts by using 50 % as the "on/off" threshold. The inset box in the reliability diagram shows the frequency of use of the various forecast probabilities and the horizontal dotted line the climatological event probability (cf. Example 5). The reliability curve (with open circles) indicates strong over-forecasting bias throughout the probability range. This seems to be a common feature at this particular location as indicated by the qualitatively similar 10-year average reliability curve (dashed line). Brier skill scores (BSS) are computed against two reference forecast systems. Of these, climatology appears to be a much stronger "no skill opponent" than persistence. The ROC curve (right) is constructed on the basis of forecast and observed probabilities leading to different potential decision thresholds and respective value pairs of H and F, as described in the text. Also ROCA and ROC_SS values are shown. The black dot represents the single value ROC from the categorical binary case of Example 5 (H=0.7; F=0.17).

6. Relating forecast verification to forecast value and forecast user's decision making

Verification measures are intended and expected to reveal the **quality** of forecasts. However, a successful forecast does not necessarily have any **value** to its final user, whereas a misleading forecast may possibly provide lots of valuable and/or useful information to another user. A forecast can be considered to exhibit value if it helps the end user to make decisions on the basis of that particular forecast, regardless of its skill. For example, forecasts of gale force winds may be (and quite often are) biased toward over-forecasting, resulting scores with low skill. Still, they may be of value to a user whose actions are economically very sensitive to strong winds.

It is highly recommended to associate with a local verification scheme features that help to **evaluate the potential economic value of the forecasts**. This is especially important in an effort to strengthen the dialogue and collaboration with customers and end users. It is quite natural that a customer would want to get some feedback on the potential economical implications of forecast information. However, the key element in this chain is the customer himself. The end forecast producer, the meteorologist, cannot have solid knowledge of the economic implications or risks of particular weather events, and even less so can the developer or producer of the background NWP guidance (like ECMWF).

Consider a decision maker who is sensitive to certain adverse weather events, for example gale force winds during a sailing event in a lake area, or occurrence of icing on a certain road network. The decision maker can then make judgements on taking some actions to prevent potential losses due to expected adverse weather. These actions would incur costs of an amount, say **C**. However, if actions were not taken and the

event would occur, the losses would amount to, say **L**. With no actions taken and no event present, the costs and losses would be nil. The example leads to the descriptive table (left-hand box) below.

Action taken	Event occurs	
	Yes	No
Yes	**C**	**C**
No	**L**	**0**

<~>

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d = n

If the end user had no forecast information available but, nevertheless, would know the climatological probability, p_{clim}, of that particular adverse weather event, he could base his decision making on the climatology and consider protective actions as follows: action is recommended if p_{clim} * L is larger than the cost of protection C, i.e.:

if $p_{clim} > C / L$ <=> action **is** recommended

if $p_{clim} < C / L$ <=> action **is not** recommended

The climatological probability of the event provides a baseline or a breaking point for the decision making. The fundamental question here is that the user should know his **Cost / Loss ratio (C/L)** upon which to establish the final decision. This, unfortunately, is quite seldom the case.

A value index (**V**) of a forecast system can be defined in a similar manner as the general form of the skill score (for more details, see [ref 6, Chapter 8] and [ref 12]):

$$\mathbf{V = (E_{ref} - E_{fc}) / (E_{ref} - E_{perf})} ,$$

where Eref refers to the expenses of using a reference forecast like climatology or persistence, E_{fc} to the expenses of the forecast system under evaluation, and E_{perf} to expenses of a perfect forecast system. V has the value = 1 for a perfect system and equals = 0 when the forecast system has the same value as the reference (like the skill score). By linking the cell count notation of the above table's right-hand side with the left-hand side theoretical costs and losses, and considering a situation there were no guidance whatsoever available, i.e. E_{ref} were defined to take protective action (incurring costs C) in every case (n), we would have:

$$\mathbf{E_{ref} = nC}$$

$$\mathbf{E_{fc} = aC + bC + cL + d0}$$

$$\mathbf{E_{perf} = (a+c)\,C}$$

The value index would then result in:

$$\mathbf{V = \{ (c + d) - ((c / (C/L)) \} / (b + d)}$$

Such an index would be easy to compute for whatever 2*2 situation, provided again, that the user-defined cost/loss ratio is known. Index V varies typically between zero and one and is **highly dependent** on C/L.

The cost/loss considerations provide a link between the end users' forecast value and standard verification measures. It was mentioned in the previous chapter that for a deterministic forecast, the ROC-based skill score ROC_SS translates to the KSS (= H - F). It can also be shown ([ref 6, Chapter 8]) that the KSS produces the maximum attainable value index (V_{max} = H - F). This would indicate that the maximum economic value is closely related to forecast skill and that skill scores ROC_SS and KSS can be related to, and interpreted as, measures of potential forecast value in addition to forecast quality. The economic value

and cost/loss discussion can be extended to probabilistic forecasts. The verification web pages of ECMWF ([web 8]) provide more insight into this area. The MSs are encouraged to apply such methodology, and what is introduced here, in their local applications in support to what is being done at ECMWF.

To summarize (including the general guidelines), and indicating minimum and optimum requirements, it is proposed to:

- minimum proposal: initiate economic value and Cost/Loss experimentation studies "inhouse" and with local forecast end users
- optimum proposal: elaborate comprehensive studies linking actual verification results (covering e.g. KSS and/or ROC_SS) with true C/L figures, including computation of value index V

7. Other issues

In addition to what has been presented heretofore, the MSs are welcome to implement and report upon **any verification related issues**. The previous text has covered mostly objective verification methods. It is stated in the annual request letter to MSs to report also on local **subjective verification** methods and results. Such activities are warmly encouraged. These are usually visual, so-called "eyeball", verifications by utilizing some kind of classification or scoring schemes. Since this has been a continuing practice for a long time in some MSs, it's continuation is essential to extend **trend evaluation** to the foreseeable future.

Another area where objective or statistical verification measures may not necessarily be applicable is **case studies**, object- or event-oriented investigations of limited time and/or spatial coverage. Such studies are occasionally reported in the "Green Book" and can provide to ECMWF and other MSs alike valuable and detailed information on local model behaviour.

Final word: Weather forecast verification is a multi-faceted act (read "art") of numerous methods and measures. Their implementation and inclusion into everyday real-time practice, seamlessly attached to the operational forecasting environment is one fundamental way to improve weather forecasts and services. Active feedback and reporting of related activities and innovations will serve the whole meteorological community.

References

Literature

[ref 1] Bougeault, P., 2003. WGNE recommendations on verification methods for numerical prediction of weather elements and severe weather events (CAS/JSC WGNE Report No. 18)

[ref 2] Proceedings, Making Verification More Meaningful (Boulder, 30 July - 1 August 2002)

[ref 3] Proceedings, SRNWP Mesoscale Verification Workshop (De Bilt, 2001)

[ref 4] Proceedings, WMO/WWRP International Conference on Quantitative Precipitation Forecasting (Vols. 1 and 2, Reading, 2 - 6 September 2002)

[ref 5] Wilks, D.S., 1995. Satistical Methods in the Atmospheric Sciences: An Introduction (Chapter 7: Forecast Verification) (Academic Press)

[ref 6] Jolliffe, I.T. and D.B. Stephenson, 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Sciences (Wiley)

[ref 7] Stanski, H.R., L.J. Wilson and W.R. Burrows, 1989. Survey of Common Verification Methods in Meteorology (WMO Research Report No. 89-5)

[ref 8] Cherubini, T., A. Ghelli and F. Lalaurette, 2001. Verification of precipitation forecasts over the Alpine region using a high density observing network (ECMWF Tech. Mem., 340, 18pp)

[ref 9] Murphy, A.H. and R.L. Winkler, 1987. A General Framework for Forecast Verification (Mon. Wea. Rev., 115, 1330-1338)

[ref 10] Stephenson, D.B., 2000. Use of the "Odds Ratio" for Diagnosing Forecast Skill (Weather and Forecasting, 15, 221-232)

[ref 11] Grazzini, F and A. Persson, 2003: User Guide to ECMWF Forecast Products (ECMWF Met. Bull., M3.2)

[ref 12] Thornes, J.E. and D.B. Stephenson, 2001. How to judge the quality and value of weather forecast products (Meteorol. Appls., 8, 307-314)

Websites

[web 1] http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html
- WMO/WWRP Working Group on Verification website

[web 2] http://www.rap.ucar.edu/research/verification/ver_wkshp1.html
- Making Verification More Meaningful Workshop (Boulder, 2002)

[web 3] http://www.chmi.cz/meteo/ov/wmo
- WMO/WWRP Workshop on the Verification of QPF (Prague, 2001)

[web 4] http://www.sec.noaa.gov/forecast_verification/verif_glossary.html
- NOAA/SEC Glossary of verification terms

[web 5] http://isl715.nws.noaa.gov/tdl/verif
- NOAA MOS verification website

[web 6] http://wwwt.emc.ncep.noaa.gov/gmb/ens/verif.html
- NOAA EPS Verification website

[web 7] http://www.wmo.ch/web/www/DPS/SVS-for-LRF.html
- WMO/CBS Standardised Verification System for Long-Range Forecasts

[web 8] http://www.ecmwf.int/products/forecasts/d/charts/verification/eps
- Verification of ECMWF Ensemble Prediction System

[web 9] http://www.ecmwf.int/products/forecasts/guide
User Guide to ECMWF Forecast Products