Suggested methods for the verification of precipitation forecasts against high resolution limited area observations

JWGFVR

Second Draft, November 2013

1. Introduction and purpose of the document

This document describes suggested verification methods for the verification of NWP-based quantitative precipitation forecasts (QPF), including ensemble forecasts. It is a follow-on to the 2009 WMO document "Recommendations for the Verification and Intercomparison of QPFs and PQPFs from Operational NWP Models", which focused primarily on global models. This document is intended both to extend the suggested methodology to higher resolution models, and to update that document, since several new verification methods have been published since the 2009 document was completed.

This document was requested by the WMO Working Group on Numerical Experimentation (WGNE). The document is needed:

1. In order to update the verification procedures used by WGNE members to evaluate model precipitation forecasts, including high resolution model output and ensemble forecasts. The verification may use, but is not limited to, radar precipitation estimates and/or high resolution non-standard gauge datasets that are available only to the National Meteorological Service of the country of origin.
2. To provide advice on which of the recently published scores are most appropriate for use in this application.
3. To update the advice provided in the 2009 WMO document on precipitation verification.

The user community for the recommended verification methods is assumed to be mainly modelers; recommendations are made to enhance the diagnostic capabilities of the verification methods. In addition to diagnostic tools, it is assumed that intercomparison of summary results among the members of WGNE may also be of interest.

Where measures are recommended that have been described in the 2009 document, the reader is referred to that document for details of the method itself. Proposed measures which are new are discussed more fully in this document, with references to relevant publications.

2. Recommendations

2.1 What should be verified

It is recommended that 6h precipitation accumulation be the primary temporal resolution of the verification. If higher temporal resolution observations are available (1h, 3h) then these forecasts could also be verified.

It is recommended that the same thresholds of precipitation amount be verified as proposed in JWGFVR (2009), 1, 2, 5, 10, 20, 50 mm per 6h, and higher if occurrences are reported. Verification should not be carried out for thresholds where there are fewer than 10 occurrences in the dataset.

2.2 Data processing recommendations

The primary verification should be done with respect to station observations. Observations should be matched to model output using the nearest grid point to avoid the smoothing effects inherent in interpolation methods.

Verification against gridded observations should also be done where possible. The analysis procedure and data quality control must be free of information from models. The further development and use of model-independent high resolution combined radar-gauge analyses is encouraged. Verification against gridded observations should only be carried out at those grid points supported by observations.

The verification data and results should be stratified by:
(a) lead time (6h, 12h, etc.)
(b) season (winter, spring, summer, autumn, as defined by 3-month periods, DJF, MAM, JJF, SON)
(c) region (E.g., tropics, northern extratropics, polar regions, etc, for global models and datasets, or specific regions defined by the availability of high resolution observation datasets)
(d) observed rainfall intensity threshold

Verification with respect to quantile thresholds is encouraged in addition to sample stratification. The quantiles can be determined from the verification sample, or obtained from long term climatology of the verification location if available (preferred but requires external data). This is especially useful for more extreme events, where low base rates and stratification severely reduces the number of occurrences. Quantile thresholds can be used instead of seasonal and regional stratification (where the quantiles are referenced to seasonal and regional climatology), and can be converted back to the physical thresholds for specific stations as needed for reporting purposes. Quantile thresholds also make intercomparison of results much more valid, and avoid the “false skill” issues that arise with climatology-sensitive scores such as the ROC area and the Brier and Continuous Ranked Probability Skill Scores. Suggested quantile thresholds are: 50% (median), >75%, >80%, >90%, >95% and <25%, <20%, <10%, <5%. Especially in dry climates, the low thresholds will probably all correspond to 0 precipitation.

All aggregate verification scores should be accompanied by 95% confidence intervals. Reporting of the median and inter-quartile range for each score is highly desirable.

2.3 Recommended measures

The following is the list of measures recommended for use in the WGNE verification. These are the common measures that all centers should compute for possible intercomparison.
For deterministic model forecasts:

- Equitable threat score (ETS)
- Extremal dependency index (EDI)
- Fractions skill score (FSS) (where gridded observations are available)

For probabilistic forecasts interpreted from ensembles, or by statistical post-processing

- Brier skill score (and components if desired)
- ROC area
- Continuous ranked probability skill score

Other measures that are easily computed and useful for additional diagnosis are: probability of detection (hit rate), false alarm ratio, frequency bias, all easily computed from the contingency table that is used for the ETS and EDI. Also, while the Brier skill score and ROC area give summary measures of skill and discrimination respectively, much more diagnostic information can be obtained from the components of the Brier skill score, reliability and resolution, and by plotting the reliability diagram and the ROC curve.

The new SEEPS score was considered for deterministic model verification, but may not be suitable for application to high resolution non-standard datasets since it requires the use of long-term climatology, which is unlikely to be available for special datasets. Further independent testing of this score is needed to establish its suitability as a standard in international intercomparison experiments.

3. Discussion

3.1 Diagnostics

It is advisable in verification projects to examine the data before computing the scores. This can be done by means of scatter plots and quantile-quantile plots, or, for categorical data, by histograms. Histograms and Q-Q plots of the observations and the corresponding model forecasts give a quick visual indication of similarities and differences in the forecast and observed distribution of precipitation. Scatterplots or histograms are also quick ways of checking the dataset for obvious errors.

For high resolution forecasts, it becomes more useful to know whether the improved forecast resolution is leading to reasonable simulation of finer spatial scale structure in the observations. One way to diagnose this is by means of a spatial spectral analysis, carried out separately for the model forecasts and for the gridded observations. An example of such an analysis is shown in Figure 1, for various versions of the UK model and for radar-based observations over the UK. This example supports the common finding that model forecasts lose definition at resolutions higher than about 6 grid lengths.

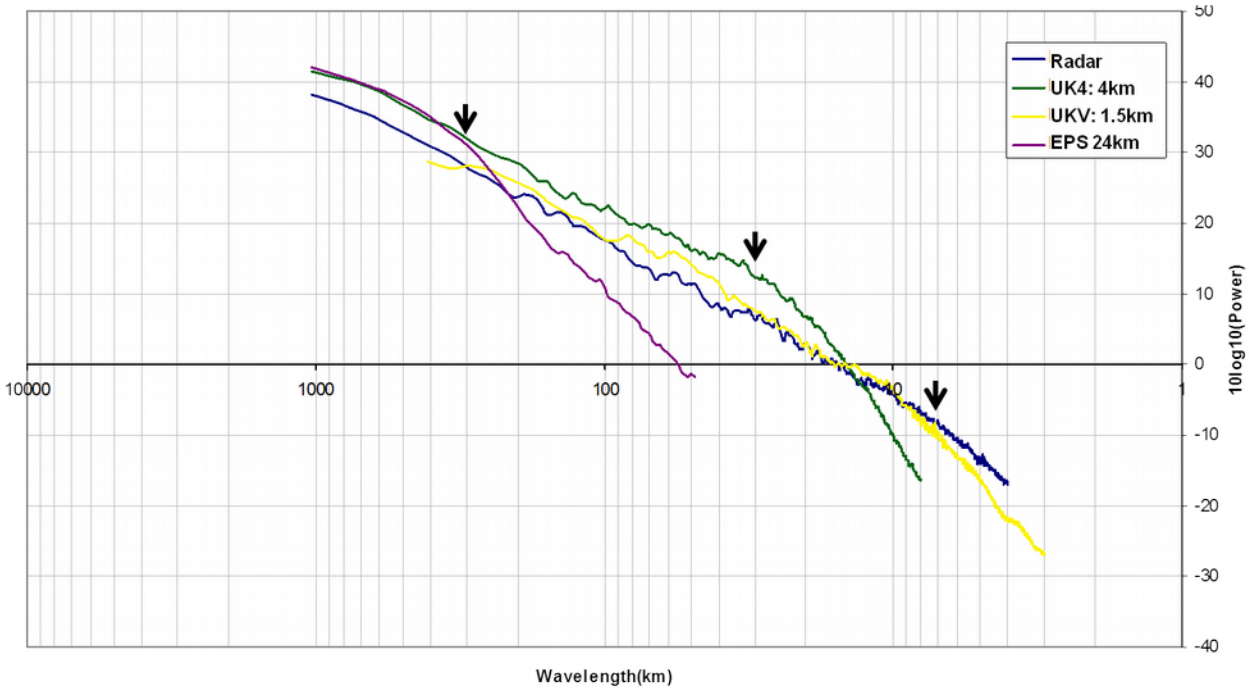

Figure 1. An example of a spectral analysis of precipitation forecasts from the UK models, MOGREPS and radar observations (Courtesy Clive Pierce via Brian Golding)

3.2 Discussion of scores recommended for deterministic forecasts

Equitable threat score (ETS)

This score is chosen primarily because it is a "legacy" score, having been widely used for summary verification for many years. It measures the accuracy of a categorical forecast, and takes into account variations in the "difficulty" of the forecast by allowing for the number of forecasts correct by chance. Since it is equitable, it is more robust for comparison of results based on different datasets. In recent years, it has been shown that this score is somewhat sensitive to the base rate (frequency of occurrence of the event in the verification sample), and it becomes insensitive to changes in accuracy for low base rates (rare events). For this reason, it is not so useful a score for rarer precipitation events. Details of the ETS are described in JWGFVR (2009) and will not be repeated here.

Extremal dependency index (EDI)

Since 2008, four new contingency table scores have been proposed to avoid the problem of base rate dependency mentioned above. Called extreme dependency score (EDS), stable extreme dependency score (SEDS), extremal dependency index (EDI) and symmetric extremal dependency index (SEDI), these four measures are all discussed and compared in Ferro and Stephenson (2011). In that paper, it is mentioned that the EDS and to a lesser extent the SEDS are still sensitive to the base rate. It has also been pointed out that the EDS does not consider false alarms, and so can be improved by overforecasting the event of interest. These shortcomings lead us to eliminate these two from further consideration. The remaining two, EDI and SEDI are both insensitive to base rate; they also are both

formulated to use the hit rate(H) and false alarm rate(F) only, which links these scores to the Pierce Skill score (H-F) and the ROC curve frequently used in verification of probability forecasts. The SEDI is symmetric in the sense that relabeling the events as non-events and non-events as events does not change the score value. The symmetry property is less important for verification of real forecasts, but may be of importance for determining the potential accuracy of a system (Ferro and Stephenson 2011). We therefore recommend the use of the EDI especially for low base-rate thresholds, but it will give a good comparative estimate of accuracy for all thresholds. The score is given by:

$$EDI = \frac{\log F - \log H}{\log F + \log H}$$

Figure 2 shows this score (and others) applied to some ECMWF precipitation forecasts. In the figure, the EDS and SEDS show lower scores for higher base rates and higher scores for lower base rates, consistent with the example in Ferro and Stephenson (2011). In the case of the EDI and SEDI, one can be sure that the differences shown for the different thresholds are due to variations in accuracy only, as measured by H and F. This figure also clearly shows the tendency of both the ETS and Pierce skill score to go to 0 for low base rates.

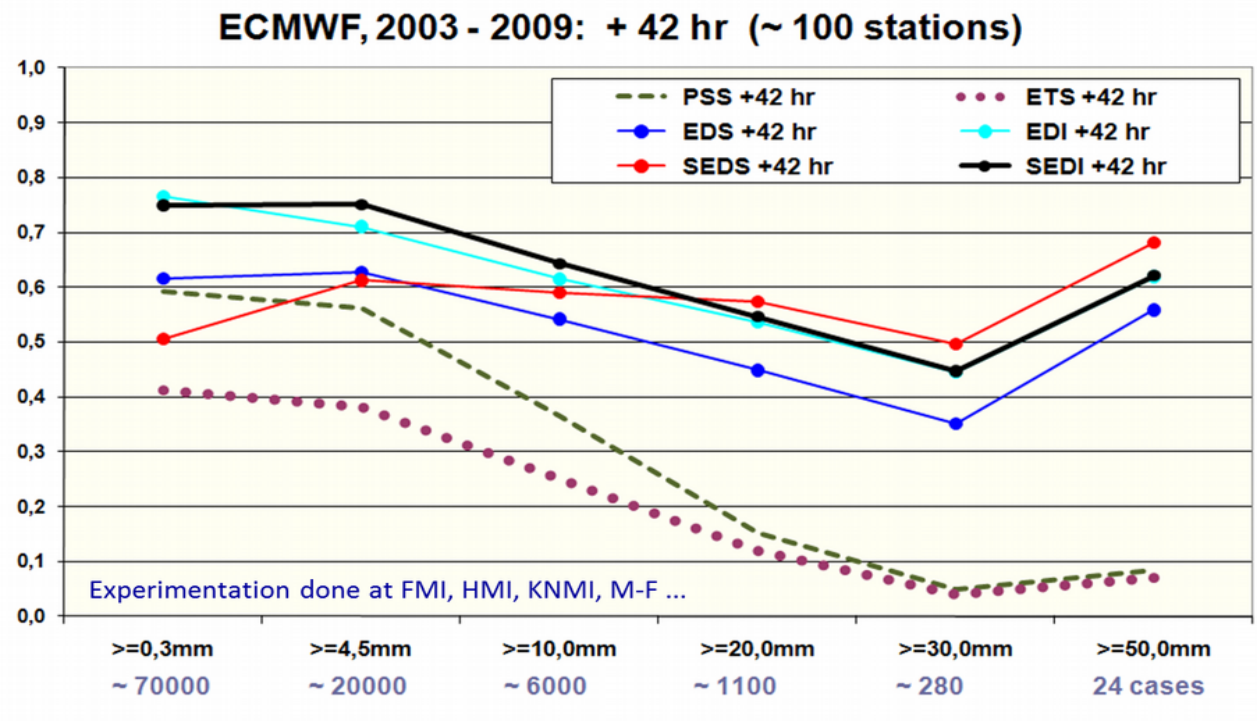

Figure 2. Comparison of EDS, SEDS, EDI, SEDI, Pierce Skill Score (PSS) and ETS for a large sample of ECMWF 42 h precipitation forecasts, as a function of threshold.

A score for spatially-oriented verification – the fractions skill score (FSS)

There has been considerable research activity devoted to the development of measures which are sensitive to the relationship in space between forecast and observations. For high resolution forecasts especially, useful diagnostic information can be obtained from these methods and one of them is recommended for implementation by WGNE members.

The fractions skill score is one of the more frequently used spatial scores, since it was first proposed in 2006. It has been included in the operational suite at the UK Met Office. Admittedly, the selection of this score is based to some extent on the fact that it is one of the easiest of the spatial methods to implement, and that it can be used over relatively small domains. All of the spatial methods proposed over the last 15 years are still under test, and future studies might reveal a more suitable alternative to the FSS. For example, the forthcoming Mesoscale Verification Intercomparison in Complex Terrain (MesoVICT) project will help focus and refine our understanding of spatial methods' unique characteristics for specific applications.

The FSS is a "neighbourhood" score in the sense that it is designed to give credit to correct forecasts of the fraction of precipitating grid boxes in the neighbourhood of a grid point, rather than requiring an exact match in space between forecast and observation. To implement the score, a precipitation threshold is determined as for the other scores recommended here; then a square neighbourhood is decided, starting with the smallest, 1x1 grid box, 3x3, 5x5, etc., right up to the largest neighbourhood, the whole domain. For each neighbourhood, the fraction of observed and forecast grid boxes with precipitation surpassing the threshold is computed. The score is a variant of the Brier Skill Score involving the mean square difference between the observed and forecast fractions, P_o and P_f, for each center point and each neighbourhood.

$$FSS = 1 - \frac{\frac{1}{N}\sum_{N}(P_f - P_o)^2}{\frac{1}{N}\left[\sum_{N}P_f^2 + \sum_{N}P_o^2\right]}$$

The score is recomputed over the whole domain for each neighbourhood size, then plotted. Of course the value of the score increases as the neighbourhood size increases. Once the curve is obtained, the range of scales with useful skill can be estimated, and the overall frequency bias can also be obtained. The FSS should be computed over a sufficiently large set of neighbourhood sizes in order to diagnose the subset of scales for which the forecasts are skillful. If the FSS is reported for a particular neighbourhood size, then the results can be compared among models only if the neighbourhood sizes are matched. Further information on the score and its implementation and interpretation can be obtained in Roberts and Lean (2008) and Mittermaier et al (2013).

Other measures

The three measures described above are considered to be most useful for diagnosis of QPF for the reasons given, but, once the data processing has been done, a few other scores can be obtained with a

trivial amount of extra work. For example, a contingency table needs to be constructed for the ETS; once that is done, several other contingency table scores can be obtained with a single extra line of code each. Of these, the hit rate and false alarm ratio (considered together), the frequency bias, and the Pierce skill score would provide the most useful additional information.

3.3 Discussion of scores recommended for probability forecasts

Brier skill score

This score is recommended because it gives a good basic measure of the skill of a probability forecast, with respect (usually) to climatology. It is recommended for this application that the sample climatology be used as the "standard" of comparison, but the reference climatology should be computed separately for each location rather than pooled over all locations. This takes account of the "false skill" effect reported in Hamill and Juras (2006). Details of the BSS and its decomposition into resolution and reliability components are described in JWGFVR (2009).

ROC and ROC area (ROCA)

The relative operating characteristic curve (ROC) describes the performance of a probability forecast system in terms of its ability to discriminate situations leading to the occurrence of the event from those which don't. Like the EDI, it uses the hit rate and false alarm rate in its construction, and so is related to the EDI and other scores which depend only on the H and F, such as the Pierce skill score. As shown by Richardson (2000), it can also be used to indicate the potential value of a forecast.

The ROCA, the area under the ROC curve is the most often used summary measure of discrimination. Its calculation has been problematic in the past, however. There are two "correct" ways to compute the ROCA, the empirical and bi-normal model. An example of the empirical method is shown below. This method is the most fundamental and accurate method, since it involves no assumptions. But, to avoid errors, it is absolutely necessary that the H and F be plotted on the diagram for every possible (allowable) forecast probability value. For example if an ensemble system has 20 members, then the discrete interpretation of probabilities means that there would be 21 possible forecast probability values, 0, 0.05, 0.10 …0.95, 1.0. The sample of forecasts would be binned into those 21 bins, the H and F computed for each, and plotted. Triangulation can be used to determine the area under the curve. In the example shown, there are only 6 possible probability forecasts, and only 15 cases in the sample. The empirical ROC computation should be used for small samples, but can be used for any size sample.

The common practice of binning the data into 10 bins and using triangulation to obtain the ROCA is incorrect unless there are only 11 allowable forecast values (including 0). If the data is binned into a number of bins which is smaller than the allowable forecast probability values, then the binormal model should be used to fit the curve. The assumptions associated with the bi-normal model are weak and are usually satisfied, except in situations where the distribution of forecasts is bimodal. The process of fitting the curve is to take the binned data, and fit a straight line after transformation to standard normal deviate space. After transformation back to H and F, a smooth curve results; this usually fits the data

quite well. The statistics package "R" contains a ROC fitting program, which also offers confidence intervals on the points along the curve.

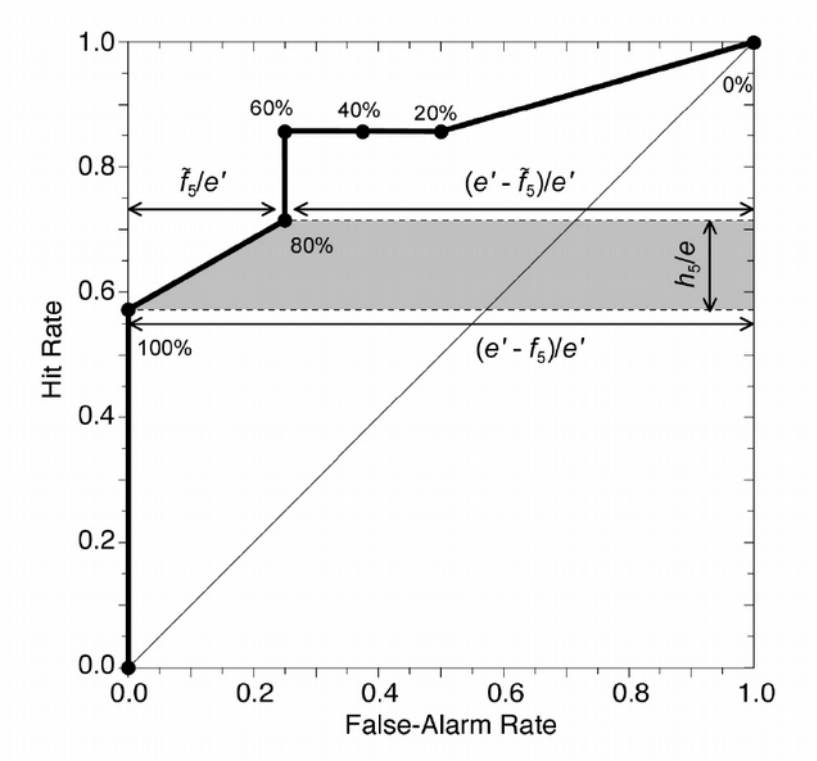

TABLE 4. FORECASTS FROM TABLE 1 SORTED BY DECREASING FORECAST PROBABILITY, WHERE THERE ARE TIED FORECAST PROBABILITIES FOR DIFFERENT CASES

Forecast	Event (1)/ non-event (0)	Probability (%)
1984	1	100.0
1985	1	100.0
1994	1	100.0
1995	1	100.0
1988	1	80.0
1981	0	80.0
1982	0	80.0
1986	1	60.0
1987	0	40.0
1991	0	20.0
1989	1	0.0
1983	0	0.0
1990	0	0.0
1992	0	0.0
1993	0	0.0

Figure 3. An example of the empirical computation method for the ROCA. (From Mason and Graham, 2002)

The empirical method is suitable for all sample sizes; the fitting method should only be applied if the bins have enough cases in them, say, at least 10 cases per bin. Bins can be chosen so that they do have enough cases, for example by using more bins at lower probabilities for low base rate situations. Four bins at least are needed.

More information on the ROC and ROCA and its interpretation is contained in JWGFVR (2009) and on the JWGFVR website.

Continuous ranked probability skill score (CRPSS)

This score is recommended because it evaluates the full probability distribution as interpreted from the ensemble, rather than just probabilities of specific events. The skill score version is chosen because it is more robust with respect to comparison over different samples. The reference forecast normally should be climatology, based on the sample, and also computed for each separate verification location. The definition and further details about the CRPSS are contained in JWGFVR (2009).

Other diagnostic tools

The three recommended measures for probability forecasts measure among them different attributes of the forecasts. However, they also summarize the information into a single number. For additional diagnostic information, the Brier Score can be decomposed into its components (reliability, resolution, and uncertainty), and those can be comparatively evaluated. It is also useful to plot both the reliability diagram and the ROC curve for more detailed diagnostic information. If the ETS and EDI are computed on the same sample as the ROCA, then it is possible to plot a point on the ROC diagram, corresponding to the H and F for the deterministic (categorical) forecast. This allows a simple comparison of the deterministic forecast and the EPS.

4. References

Ferro C.A.T., and D.B. Stephenson, 2011: Extremal Dependence Indices: improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting*, **26**, 699-713.

Hamill, T.M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. Royal Met. Soc.*, **132**, 2905-2923.

JWGFVR, 2009: Recommendations for the Verification and Intercomparison of QPFs and PQPFs from Operational NWP Models, WMO TD No. 1485, WWRP 2009-1.

Mason, S.J. and N. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.* **128**, 2145–2166.

Mittermaier, M., N. Roberts and S. Thompson, 2013: A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorol. Appl***. 20,** 176-186.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Met. Soc.*, **126**, 649-667.

Roberts, N.M. and H.W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78-97.