WG 3: Verification and applications of ensemble forecasts

Participants: *Ken Mylne (Chair), Martin Leutbecher (secretary), Magdalena A. Balmaseda, Paco Doblas-Reyes, Lizzie Froude, Jose A. Garcia-Moya, Anna Ghelli, Renate Hagedorn, Edit Hagel, Florian Pappenberger, Fernando Prates, Anders Persson, Cristina Primo, Kamal Puri, Thomas Schumann, Olivier Talagrand, Jutta Thielen, Helen Titley*

First, the discussions of the working group on the Verification of Ensemble Forecasts and the Application of Ensemble forecasts are summarised in Sections 1 and 2, respectively. This is followed by specific recommendations for ECMWF in Section 3. Most of these recommendations are also considered to be relevant for other producers of ensemble forecasts.

1. Verification of Ensemble Forecasts

1.1. Purpose of Verification

The working group identified three different purposes of the verification:

- Objective measure for improving the EPS and guide future developments of system and funding
- Guide forecasters, service providers and users on which product(s) to trust and use. However, some participants thought that, at present, not many forecasters look at verification scores for probabilistic forecasts.
- Demonstrate usefulness of ensemble prediction systems for particular applications or various user groups (from general public to decision makers).

1.2. Statistical Significance of Results

It was recommended that confidence intervals should be provided in all verification statistics. Further research will be required in order to study methods of computing confidence intervals and significance tests and to verify the reliability of confidence intervals. Examples: Bootstrap methods, analytic methods. Latter require assumptions about the distribution. Bootstrap requires fewer assumptions but may be costly to calculate.

1.3. Accounting for the uncertainty of the verification data

There was consensus that it is necessary to account for observation errors/analysis errors in the probabilistic verification. One accepted way of doing this (for rank-histogram and probabilistic verification in general) is to add noise with the same statistics as that of the verifying data to the individual ensemble forecasts (Saetra et al). A second alternative is to use a deconvolution method (Bowler). A third method is to consider the observation as a probability distribution (Talagrand & Candille). The first method will tell us only how well we can predict an observation not how well we can predict the true state.

It was noted that it can be difficult to estimate the error characteristics of some verification data: e.g. rain gauge data.

1.4. Choice of verification measures

There are many different verification measures in use, and the group did not attempt to review them all. Discussion centred around how to communicate performance effectively to decision-makers and users who may not be specialist in the details of the science.

- The choice of the verification will reflect its purpose (see above).
- There was general consensus that classic upper air verification has its value in providing guidance concerning the general accuracy and reliability of a (probabilistic) prediction system. It should be complemented by verification of surface weather variables which are particularly relevant to users and for short-range ensemble prediction.
- The working group agreed that *several* measures are required to examine different aspects of the Probability Density Function (PDF). For instance, the rank histogram, on the one hand, and the reliability component of the Brier score with respect to a threshold, on the other hand, measure different aspects of reliability. A single summary measure (although desirable from a management perspective) was considered inadequate.
- It was noted that reliability is a statistical property, and global reliability, as estimated by a particular measure over a given set of realizations of an EPS, may always result from mutual compensation between individually unreliable subsets (Example: flatness of a rank-histogram is a necessary but not a sufficient condition for the statistical reliability of an EPS).
- The decomposition of scores to better understand results was encouraged where appropriate for the audience (e.g. reliability and resolution component of Brier-type scores)
- Concerning recommendations for system upgrades, selective use of particular scores was discouraged as this may encourage the selection of the favourable subsets. In other words, the impact of system upgrades (e-suite versus o-suite) should always be documented by the same standard set of scores (further discussion is required which ones these should be).
- Use of Reliability Diagrams (including sharpness of the a posteriori calibrated probabilities) was considered to be the easiest way to communicate probabilistic verification to non-specialist audiences.
- There is need for care in the use of the Relative Operating Characteristic, and the area under the ROC in particular, as a verification measure. It is particularly difficult to explain to users and non-specialists. Value of ROC area is very sensitive to how it is calculated and to small changes in performance. Use of confidence measures would help.
- Probabilistic scores need to be complemented by more general model validation: (ability to simulate climatological mean and variance, …)

1.5. Avoidance of False Skill (added by Chair)

It was noted, but not discussed in detail due to time constraints, that without due care verification can indicate more skill in forecast systems than is warranted due to climatological differences between locations included in the verification set - the base rate effect. This was illustrated by Tom Hamill's presentation in the workshop. It was noted that ECMWF has already taken steps to avoid this in some of the verification presented at the workshop. Two methods are proposed to minimise the effect:

- Define events for probabilistic verification as percentiles of the climatological distribution rather than absolute values [eg. p(T>90th percentile) or p(T> 1 s.d. above normal) but not p(T>5 Celsius)]. (This method is proposed by Hamill and Juras, and was used by ECMWF in some verification presented at the workshop.)
- Proposed by Hamill in his presentation: group sites according to their climatological frequency of the event (eg group all sites for which 5mm/24h occurs with climatological frequencies of 0-5%, 6-15%, 16-25%, etc). Estimate the geographical area for which each climatological category is

representative. Calculate verification statistics for each group of sites, and then average results weighting the values according to the proportion of geographical area represented by each group.

1.6. Verification of rare/extreme events

Some discussion focussed on whether a different methodology needs to be adopted for the verification of rare/extreme events (to be distinguished from high-impact events which need not be rare in the climatological sense)

- There is a problem of sample size for rare events; no statistical significance may be reached.
  - The same is true for events which occur almost always because of the symmetry of scores (event occurring / not occurring).
- The prediction of events within a finite space-time domain selected around a particular weather event or atmospheric feature (as opposed to point values) will lead to higher probabilities for some extreme events than local probabilities (example cyclone strike probabilities).
- Some extrapolation may be possible from the verification for the more moderate thresholds which will have statistically more reliable results. Use of error bars would help guide how far it is reasonable to extrapolate conclusions.
- Case studies were considered to be essential in assessing performance for extreme events. It was proposed that case studies should include cases where ensembles predict non-zero probabilities of extreme events, but which do not verify, to ensure a reliable probabilistic prediction. This should be used to complement objective statistical verification of less extreme events.
- It was noted that where extreme events do occur, the observation is very often close to the extreme of the ensemble distribution - hence users should be strongly encouraged to pay attention to low probability alerts. This has also implications for decision making (cost/loss analysis). Support from high-resolution models may strengthen the signal.
- For some cases of severe events, experience suggests that the actual predictability may be higher than implied by the EPS. Since predictability is a property of a forecast system it is of interest to quantify how the ensemble dispersion relates to the forecast accuracy of state-of-the-art deterministic models.

1.7. Aspects relevant for the verification/applications of seasonal/monthly forecasts

- The question was raised whether an effort should be made to unify verification tools used for different applications (e.g. medium-range, monthly, seasonal predictions).
- The limitations of obtaining statistically significant verification results for seasonal and longer predictions was discussed
- A member of the working group noted that the repetition of similar anomalies in subsequent years in the seasonal forecast may decrease trust of users in the useful signal in this product. It was suggested that the repetition of similar anomalies might be due to the choice of climate and its lack of accounting for the climate change trend.
- It was suggested that the climatology should include a trend; this aspect is relevant both for the verification and the communication of seasonal forecasts.

- Case studies: It was recommended to identify a priori events where the predicted PDF deviates significantly from the climatological PDF to avoid a selective verification of only the events that did verify (see discussion above on extreme events).
- Discussion was held around whether seasonal forecasts should be issued in areas or seasons with little or no skill. It is always important to communicate information on the level of skill.
    - In cases of no skill it is preferable that no forecast should be issued, but the user could be provided with the climatology.
    - Where there is some skill and forecasts are issued, it was recommended that forecasts should be issued consistently, including those occasions when there is no strong signal. In this case the forecast should revert to climatology together with the information that there is no strong or useful signal in the seasonal forecast.
    - In summary, issue forecasts where there is skill but no signal, but NOT where there is signal but no skill.
- It was mentioned that the consistency of subsequent forecasts can increase trust of users in the product (as an example the monthly forecast was mentioned).
- It was briefly discussed whether the communication of seasonal predictions in the form of anomalies with respect to the recent *N* years (with N being some number up to 10) may be more useful and/or easier to interpret for some users than anomalies with respect to a longer term climate which includes periods beyond the personal memory of the users. This was considered particularly useful where the climate has undergone significant change in recent decades.

1.8. Benchmark(s)

Some “fair comparison” of EPS with probability distributions built from the High-Resolution deterministic forecast should be examined. A simple example of such probability distributions are Gaussian distributions centred on the deterministic forecast with a standard deviation depending on forecast range.

1.9. Educational Aspects of Verification

- It was felt that more education/explanation of the meaning of scores and changes of the scores was required. A short guide to the meaning/usefulness of different verification measures and their appropriateness for different applications would be useful.
- When results of EPS verification are presented, a range of scores is required.
- The relative meaning of different scores must be explained as the apparent meaning of some results may be counterintuitive (a well known example for deterministic forecasts is the reduction of RMS error when the activity of the model is reduced and vice versa).

2. Applications of Ensemble Forecasts

2.1. Communication of probabilistic forecasts

- A WMO guide on the communication of probabilistic forecasts will be published soon. (*It is not published on web yet – please contact ken.mylne@metoffice.gov.uk for update*)
- Use and limitation of Ensemble Mean. There was some disagreement on the value of the ensemble mean. Some experience suggested that it was useful in introducing use of ensembles into predominantly deterministic environments. However, there was considerable concern about the limitations of the mean, in particular its inability ever to predict extreme events and the view was

also expressed that the ensemble mean should never be used. Where used it should always be accompanied by probabilities of extreme or high-impact events.

2.2. Hydrological applications

- The need for re-forecasts was stressed not only for calibration but also for verification (the former should be covered by WG2)
- A limited sample size is an issue for flood forecasting. For instance, the definition of a climatological PDF for streamflow from a catchment poses a problem. Thus, it is difficult to evaluate skill scores or objectively compare different ensemble configurations
- There may be particular verification difficulties arising from the effects of river control measures and changing river profiles which impact the consistency of the observations.

2.3. Recommendations concerning design and testing of forecast system

- Based on predictability theory, medium-range and later range forecasts should be issued in probabilistic form. Therefore, the prime aim of ECMWF should be to provide the tools to predict a reliable and sharp PDF of the atmospheric state rather than just a single deterministic high-resolution forecast. The PDF should be based on all available information, i.e. the EPS and the high-resolution deterministic forecast. Consequently, the EPS should be given the same level of attention as the deterministic forecast system. Research on how to best combine high-resolution deterministic forecast and EPS should be continued at ECMWF. It was, however, also stated that a user/application-specific combination may be superior to a generic combination. In such cases, the production of an optimally combined PDF would fall into the responsibility of member states or individual end-users.
- The EPS should therefore be afforded:
  - Sufficient computer resource to allow optimal model performance
  - tuning of the forecast model/physical parameterisations
  - duration of experimental suites
- It was felt that it should be further investigated whether there is a benefit of going from 62 vertical levels to 91 in the EPS (eg benefit of improved stratospheric resolution on medium-range forecast).

3. Summary of Recommendations for ECMWF and other EPS Producers

3.1. Recommendations on Verification

- Confidence intervals should be provided in all verification statistics. Some research is required in the best techniques for estimating confidence intervals, but there is already some useful work in the literature.
- Methods for accounting for observation or analysis error should be applied in calculation of verification results. Several methods are proposed in the literature.
- Methods should be used to minimise the impact of apparent “false skill” in verification results caused by differences in climatological frequency of events (base rate) at different locations.
- Several measures are required to examine different aspects of EPS performance. A single summary measure, although desirable from a management perspective, is inadequate.

- These several measures should be used in a consistent fashion. The impact of proposed system changes should be documented using the full set to give a balanced picture of the strengths and weaknesses of the change; selective use of particular scores is strongly discouraged.
- Reliability diagrams, together with sharpness diagrams of the corresponding a posteriori calibrated probabilities, provide an effective way to communicate probabilistic performance to non-specialists.
- The Relative Operating Characteristic, while useful in research, should be used with care as it is difficult to explain and the ROC area summary measure can be very sensitive to how it is calculated and is frequently mis-interpreted.
- Statistically significant verification of rare extreme events is impossible. Judicious use of confidence intervals may allow some extrapolation of results from less extreme events.
- Use of case studies for extreme events should be balanced with cases where ensembles indicated a probability of a severe event but none verified.
- *Fair comparison* should be made between EPS forecasts and the high-resolution deterministic forecast dressed with error statistics.
- Some investigation is encouraged of whether the ensemble spread reflects the true predictability of extreme events from the high-resolution deterministic model.
- A simple summary guide of commonly used probabilistic verification statistics and their interpretation is required.

3.2. Recommendations on Applications

- The prime aim of ECMWF (or other NWP systems) should be to provide tools to predict a reliable and sharp probability distribution of the future state of the atmosphere based on all available information – the needs of EPS should therefore be afforded equal attention as high-resolution deterministic forecasts, and appropriate levels of resources.
- Seasonal forecast should be issued consistently, and where there is no signal this should be clearly communicated.
- One should consider the possibility of describing the trend in the reference climatology for seasonal forecasts, to avoid issuing forecasts which are dominated by the climate change trend.