Introduction

Ensemble weather forecasting has come of age - this workshop was held on the 15th anniversary of the implementation of operational medium-range ensemble weather forecasting, back in 1992. Since that time we have seen ensemble forecasting spread to all areas of weather and climate prediction, from the shortest ranges to the longest ranges. For example, Sir Nicholas Stern acknowledged at the time of publication of his influential report on the Economics of Climate Change in 2006, that quantitative economic risk assessments of climate change are not possible without ensemble predictions of regional climate change.

At its heart, ensemble forecasts allow users to make better decisions than they might make with just a single forecast. For example, a wind turbine cannot function if the wind speed is too high; as such it is critically important for deciding how much power to contract to produce from a given turbine, the probability that the wind exceeds this threshold.

Empirically we know that current operational ensemble forecast systems have considerable skill in forecasting probabilities of relevant weather and climate events. However, we are still a long way from producing a good theory for describing the uncertainties that arise when predicting weather or climate. Different groups have developed different techniques, and through the THORPEX data archives we are able to compare and contrast these techniques.

This workshop provides an excellent overview of where we are in ensemble forecasting, from the short range to extended range. Many of the world's leading ensemble prediction systems are described. Papers on validation and specific applications are also included.

Fifteen years ago, ensemble prediction was a tentative step forward. We were all fully aware that if the probability products weren't used, the experiment would be curtailed. Now it is clear that ensemble prediction has become an established technique and all the major forecast centres around the world have developed an ensemble forecast system. However, there is still a long way to go. From a practical point of view, how do we express forecast uncertainty to the public? From a theoretical point of view, how do we formulate model uncertainty in a rigorous way? These are examples of questions which lie at the extremes of a spectrum of questions that will characterise research on ensemble forecasting in the coming 15 years, research that will attract talented individuals from a range of backgrounds.

Half of the three day meeting was devoted to lectures by invited speakers and the remaining time was allocated to discussions in working groups and a plenary session. The topics for the working groups were "Representing initial and model uncertainties", "Methodologies for downscaling and calibration" and "Verification and applications of ensemble forecasts". The discussions and recommendations are summarized in three reports published in these proceedings. The contributions to the workshop have also been posted on the ECMWF web site http://www.ecmwf.int/publications/.

ECMWF thank all the participants for contributing to a successful and stimulating workshop.

WG1: Representing initial and model uncertainties

The Working Group (WG) on representing initial condition and model uncertainty met at ECMWF on November 8-9, 2007. There were 23 participants, and Roberto Buizza and Ben Kirtman co-chaired the WG. The following summarizes the WG discussions and concludes with some ECMWF specific recommendations that are motivated by the consensus opinion that emerged from the deliberations.

The WG deliberations began with a detailed discussion of a series of questions:

1 What is the main contributor to forecast uncertainty? Initial condition uncertainty or model uncertainty?

The WG discussed whether initial condition uncertainty can be separated from model uncertainty. It was immediately recognized that model uncertainty makes significant contribution to initial condition uncertainty. Nevertheless, it is possible to diagnose the contribution to forecast error associated with model error by assuming a perfect model approach initializing forecasts using the Ensemble Kalman Filter (EnKF), and to a lesser degree using the ensemble 4-dimensions variational approach.

The WG had extensive discussions regarding the fact the model error and uncertainty needs to be diagnosed at the process level. Indeed, the WG felt that it was important to distinguish between model error associated with limitation in horizontal and vertical resolution versus errors in the parameterization of the physical processes. The WG also was careful to acknowledge that resolution and parameterization are closely linked.

The WG acknowledged the fact that there is still no agreement on whether one of techniques used to simulate initial uncertainty is superior to the others, and recognized the value of studies designed to compare ensembles based on different methodologies. But overall, the WG concluded that given the current status of ensemble prediction systems, research priority should be given to improving the simulation of model uncertainty and reducing model errors/biases.

2 Are ensemble prediction systems sampling initial condition uncertainty properly?

The WG consensus is that there is a need to improve the sampling of initial condition uncertainty. The WG discussed the fact that current techniques (e.g. singular vectors and bred vectors) for sampling initial condition uncertainty selectively probe that uncertainty (i.e., they capture the fastest growing perturbations), and that there is still no agreement on whether such a 'selective sampling' approach is needed. The WG pointed out that there are serious data and modeling issues associated with sampling land surface initial condition uncertainty, which have received relatively little attention.

3 Are ensemble prediction systems sampling model uncertainty sufficiently?

The WG agreed that there is much work to be done in terms of improving the simulation of model uncertainty and in terms of reducing model biases. The WG recognized the current approaches to simulating model uncertainty, e.g., schemes based on stochastic perturbations added to model tendencies due to physical parameterization (i.e. the schemes operational at ECMWF and MSC) and schemes based on stochastic backscatter ideas (e.g. the ones under development at the UK Met Office, ECMWF and MSC), but also acknowledged that these approaches have limitations. The WG agreed that future efforts in simulating model errors should have linkages to the physical process and parameterizations. There was also some discussion regarding the simulation of know model errors on the largest scales with particular emphasis on the MJO. In terms of model biases/error, the WG discussed the possibility of examining the parameterization problem by including space-time coherence.

The discussion of question 3) closed with a clear consensus that the multi-model approach to ensemble prediction is an extremely pragmatic and useful mechanism for sampling model uncertainty, but it should not be used to avoid the difficult problems of improving model of model uncertainty or reducing model errors and biases.

4 How do model-component interactions (i.e., atmosphere-ocean, atmosphere-land ...) contribute to model uncertainty?

The WG concluded that there are large errors in the individual component models and in their interactions that require further documentation and understanding, and that the addition of the simulation of these interactions may increase substantially the forecast skill in some regions (e.g. the inclusion of sea ice-atmosphere interactions might improve the forecasts of temperature evolution close to the polar caps). The WG also discussed additional issues regarding the additional model complexity (i.e., biogeochemical cycles). While the WG felt that these processes might prove to be important for ensemble prediction systems, some caution was expressed. Specifically, there is the concern that these model-component interactions might add noisiness to the system that is not well understood.

The WG also discussed the utility of making weather forecasts with coupled models and agreed in principle that this is desirable.

The WG also agreed that weather forecast models should be verified in climate mode and that climate models should be verified in weather forecast mode.

5 Resolution versus ensemble size?

The WG spent some time discussing the trade off between ensemble size and ensemble member resolution. In general, the WG felt that this issue needs to be revisited with the goal of potentially increasing the resolution of the ensemble members.

The WG also thought that the relative value of a single higher-resolution forecasts versus an ensemble of lower-resolution forecasts should be re-assessed with state-of-the-art systems (no conclusions on this issue were drawn).

The WG acknowledged the value of variable resolution approaches to ensemble prediction, but pointed out that "shock" issues associated with truncating the resolution of ensemble members (e.g. from TL399 to TL255 at day-10 in the ECMWF VAREPS) need to be documented in more details.

The WG agreed on the following list of specific recommendations to ECMWF:

- More Emphasis should be given to understanding model error, and to improving the simulation of model errors on weather and seasonal time scales. Developments along this lines would benefit from a very close interaction between scientists developing paramtrization schemes and scientists developing "models of model errors"
- The impact of using of coupled models in weather forecasting systems (high-resolution, ensemble) from initial time should be documented, and if beneficial should lead to coupling from step zero
- Different ensemble configurations, e.g. with a smaller ensemble size but a higher resolution, should be investigated, taking into consideration the fact that:
    - The answer might be user/application dependent (users should be involved/consulted)
    - If a variable resolution is used, the shock due to the resolution truncation should be studied and properly documented
- The impact on the ensemble performance of raising the top of the model and of increasing the model resolution in the stratosphere should be investigated
- The TIGGE data-based is an extremely valuable data-set: ECMWF should consider to extend it to seasonal time scales
- Experiments should be designed and performed to diagnose and isolate model error (e.g. using ensemble 4-D Var) specifically by assuming there is no model error in the data assimilation system and then forecast error is due to model error.

- Work on ensemble 4-D Var data assimilation should continue:
    - 4-D Var ensemble methods should be compared with EnKF methods
    - The investigation on the use of an ensemble of perturbed 4-D Var analyses in ensemble prediction should continue
    - Collaboration with other Operational Centers on EnKF Data Assimilation should be promoted

WG2: Methodologies for downscaling and calibration

The Working Group on methodologies for downscaling and calibration met at ECMWF on Nov. 8-9, 2007. It was attended by about 15 participants, and was co-chaired by Tom Hamill and Franco Molteni. During the informal discussion on the first day, additional results relevant to the discussion were briefly presented by some of the participants. The discussion, as well as the report to the plenary session on the second day, was organised along three main themes:

i. Medium-range forecast calibration.

ii. Re-forecasts: done on the fly, or with frozen model version?

iii. Seasonal hindcasts and calibration.

A summary of the discussion topics and related recommendations are as follows.

1 Medium-range forecast calibration

A wide consensus was found within the WG on the need for re-forecast data sets in order to better exploit ensemble forecasts in the late medium-range. Regarding the added benefit of a multi-model approach, the following questions were raised:

- What is relative benefit of single model with re-forecast calibration vs. multi-model without calibration?
- Is there benefit from re-forecast data sets for multiple models, or is the benefit primarily obtained with one re-forecast?

It was agreed that it is difficult to answer such questions without appropriate experimentation, which in itself requires the availability of re-forecast sets. The role of both forecasting centres and individual users in evaluating the potential benefit of calibration for specific applications was also discussed, and the WG agreed on the following recommendations:

- Operational centres should provide reforecast data sets to member countries/users.
- Calibration efforts in the medium range should mostly be made by individual countries/users. However, some users may not have resources to perform calibration. There is thus value in having a basic set of calibrated products from operational centres.
- Operational centres/national governments should coordinate efforts to provide high-resolution gridded datasets of analyzed weather parameters to be used for calibration.
- Techniques to optimise calibration for systems with limited re-forecast availability should be further developed.

The relative advantages of dynamical versus statistical downscaling for medium-range forecasts were discussed. Depending on the relative merits of the two approaches, ECMWF may decide on whether to allocate future computing resources to higher-resolution integrations or extended re-forecasts sets. It was noted that no unequivocal answer is available from results presented so far, and the issue cannot be properly investigated without some resources being put into re-forecasts.

2 Re-forecasts: done on the fly, or with frozen model version?

The debate outlined the relative advantages and disadvantages of the two approaches.

- "**On the fly re-forecast**": continual production of re-forecasts with current operational model version.
  Advantage:
  - re-forecasts available with very latest model version including most advanced physical parametrizations.

Disadvantages:

  - Reforecasts use reanalysis from previous model version, which may reflect earlier model version's different initial-condition bias in data-sparse regions, leading to possible inconsistencies between real-time vs. training data sets
  - Re-forecast data sets change frequently throughout the years.
  - Possibly confusing for users, who must understand changes between model cycles and must continually download a changing reforecast data set.

- **"Frozen-model re-forecast"**: made from re-analyses with consistent model cycles (this may be seen as part of a regular cycle of re-analysis & re-forecast production).

  Advantage:

  - No inconsistency between real-time/training data sets. Promotes understanding of reanalysis model issues.

  Disadvantage:

  - Older model physics, possibly reduced resolution, need to maintain and operate two model versions.

Overall, the current ECMWF strategy of running on-the-fly re-forecasts for the medium-range and frozen re-forecasts for seasonal forecasting was considered as appropriate. Regarding the monthly time-scale, the cost of high-resolution on-the-fly re-forecasts was noted, but most participants also stressed the need for an appropriate temporal coverage of such datasets (with more than one decade needed for robust parameter estimation).

3 Seasonal hindcasts and calibration

A widespread agreement was found within the WG on the following statements:

- The production of hindcast sets is an established and non-controversial component of any seasonal forecasting system.
- Multi-model approaches are important and may provide additional skill given the substantial model error still exhibited by coupled models (international collaborative projects like DEMETER/ENSEMBLES, APCC/CLIPAS are particularly valuable in this respect).
- Calibrated products from seasonal forecasts should be provided (and validated) by the forecasting centres, but the availability of hindcasts for the development of tailored products from external users is essential.

With respect to current availability (e.g. from the ECMWF system), it was pointed out that additional value may be provided by saving data more frequently and increasing the number of ensemble members in the hindcasts (provided there is a demand by individual member states). On the other hand, most of participants agreed that many decades of hindcasts are necessary (~30 yr), spanning multiple cycles of ENSO, though more years may lead to observational non-stationarity induced by climate change (see WMO-CBS expert team documents on this topic) .

As for the medium-range, the relative merits of statistical vs. dynamical downscaling were discussed. Such an issue is, again, relevant to the discussion on the trade-off between increased horizontal resolution and larger hindcast datasets in future systems. Results from dynamical downscaling of seasonal forecasts available so far (see presentations in this Workshop) appear to be inconclusive about the merits of this approach, while there is growing evidence on the benefits of statistical downscaling.

The WG also noted the following:

- In most of current seasonal forecasting systems, hindcasts are typically computed in advance with a frozen system. This approach favors consistency in calibration, at the expense of a decreased real-time impact than for medium-range "on-the-fly" re-forecasts.
- When considering allocating resources to increased resolution, the relative advantages of ocean vs. atmospheric resolution should be carefully evaluated.
- In order to properly assess the value of forecasts performed with coupled models, comparisons between statistically downscaled numerical seasonal forecasts and purely statistical models (e.g., linear-inverse models) should be carried out.

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

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 is 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 is published on the WMO website at http://www.wmo.int/pages/prog/amp/pwsp/documents/TD-1422.pdf.
- 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.