World Meteorological Organization

Weather · Climate · Water

Guidelines on Ensemble Prediction Systems and Forecasting

Guidelines on Ensemble Prediction Systems and Forecasting

2012

CONTENTS

GUIDELINES ON ENSEMBLE PREDICTION SYSTEMS AND FORECASTING

1. **INTRODUCTION**

Ensemble Prediction Systems (EPS) are numerical weather prediction (NWP) systems that allow us to estimate the uncertainty in a weather forecast as well as the most likely outcome. Instead of running the NWP model once (a deterministic forecast), the model is run many times from very slightly different initial conditions. Often the model physics is also slightly perturbed, and some ensembles use more than one model within the ensemble (multi-model EPS) or the same model but with different combinations of physical parameterization schemes (multi-physics EPS). Owing to the cost of running an NWP model many times, the EPS is normally run at around half the horizontal resolution of the equivalent deterministic NWP model. The EPS normally includes a control forecast that uses the ensemble resolution model but without any perturbations to the analysis or model. The individual NWP solutions that make up the ensemble are often referred to as the ensemble members. The range of different solutions in the forecast allows us to assess the uncertainty in the forecast, and how confident we should be in a deterministic forecast. The uncertainty in a weather forecast can vary widely from day to day according to the synoptic situation, and the EPS approach provides an estimate of this day-to-day uncertainty. The EPS is designed to sample the probability distribution function (pdf) of the forecast, and is often used to produce probability forecasts – to assess the probability that certain outcomes will occur.

The present guidelines are intended to provide some general advice to forecasters and forecast providers on the effective use of EPS, and on what EPS can and cannot be expected to provide. A general working knowledge of the principles and use of NWP is assumed. For those requiring more detailed information, the *User Guide to ECMWF Forecast Products* (http://www.ecmwf.int/products/forecasts/guide/) of the European Centre for Medium-Range Weather Forecasts provides comprehensive guidance on the use of ECMWF systems including detailed advice on the use of EPS; the COMET training materials (https://www.meted.ucar.edu/training_detail.php?orderBy=&topic=15) also provide training on the use of EPS.

In general, it is strongly recommended that uncertainty should be communicated as part of every forecast. Guidance is given in the *Guidelines on Communicating Forecast Uncertainty* (PWS-18, WMO/TD-No. 1422).

Examples shown in the present guidelines are mostly taken from the United Kingdom Met Office Global and Regional Ensemble Prediction System (MOGREPS) EPS or the ECMWF EPS, but the principles described apply to any EPS.

2. **WHY SHOULD WE USE EPS?**

Numerical weather prediction systems using the latest numerical models of the atmosphere are very powerful systems to aid the forecaster in producing weather forecasts. Many models now provide a good enough representation of the weather that they can also be used to provide basic automated weather forecasts from direct model output (DMO), although in general it is recommended that some post-processing should be used to calibrate automated forecasts. Direct model output provides a better representation of some weather elements than others, for example, surface temperature is often quite well resolved (at least away from steep surface orography), whereas precipitation is often much less well resolved.

However, despite these advances, it is well known that forecasts from even the very best models can often go badly wrong. This is most obvious in forecasts several days ahead and is due to the chaotic nature of the atmosphere. We forecast the weather by starting the model from an analysis of the state of the atmosphere based on the latest observations that are taken all around the world. The model then calculates how the atmosphere will change and evolve from this initial

analysis state over the coming days. Chaos theory means that the way the atmosphere evolves is very sensitive to small errors in that initial analysis, so that a tiny error (often too small for the forecaster to even notice) can become a large error in the forecast. Even with the best observations we can never make a perfect analysis, so we cannot make perfect forecasts. This is why we run EPS (ensembles).

In an ensemble forecast we make very small changes (perturbations) to the analysis, and then re-run the model from these slightly perturbed starting conditions. If the different forecasts in the ensemble are all very similar to each other then we can be confident of our forecast, but if they all develop differently, and for example some develop a major storm while others develop a much weaker depression, then we will be much less confident. However, by looking at the proportion of the ensemble members that predict a storm, we can make an estimate of how likely the storm is.

When we look at shorter-range forecasts of 1 or 2 days ahead, the general pattern of the weather is usually much more predictable, but we can still find important differences between ensemble members when we look at the local detail of the weather which may be important to many forecast users. Also, occasionally the larger-scale evolution can be uncertain even at short range – this is most likely to happen during the development of major storms, so it is important to take account of the EPS even in short-range forecasts.

3.   TYPES OF EPS

There are three main types of EPS for use in weather forecasting – global, regional and convective-scale – and as with deterministic NWP models, they address different timescales in the forecast. These will be outlined briefly below. Within each of these categories there are many variations, such as the way in which perturbations are created and the variations in the models used within the models; however, the principles of how the ensembles are used remain the same, and these details are not covered here. (It may be noted that ensembles are also used for long-range forecasting and climate prediction. The principles are very similar, but these will not be considered in the present guidelines, which focus on forecasts of up to 15 days – the period over which it is often possible to forecast daily weather.)

3.1   Global EPS

Global EPS are normally designed and used for medium-range forecasting of 3–15 days ahead. They use global NWP models and are run at relatively low resolutions with typical grid lengths of between 30 and 70 km. Although they are primarily designed for use in the medium range, their global coverage means that they can also be used to provide short-range EPS forecasts in regions of the globe where no other EPS are available, and may be the only available option for many WMO Members. In this context they are used extensively to provide products to support the several projects of the WMO Severe Weather Forecasting Demonstration Project (SWFDP).

Forecasters using global EPS should always remember that the relatively low grid resolutions will limit the detail they can expect in the forecasts. Global EPS will often not be able to resolve details such as the full strength of wind speed in a storm.

3.2   Regional EPS

Regional or Limited Area Model (LAM) EPS use regional models over smaller areas and are focused more on the short-range forecast of 1–3 days ahead. They use higher grid length resolution than the global EPS, typically between 7 and 30 km, which allows them to forecast more local detail in the weather and also to better resolve intense weather systems. Nevertheless, the forecaster should remember the limitations of resolution; for example, a regional EPS should not be expected to predict details of small-scale systems such as thunderstorms.

A regional EPS has to take its lateral boundary conditions (the weather systems moving into the area from outside the domain) from a global EPS. Some regional EPS use a high-resolution regional analysis and calculate corresponding high-resolution perturbations, but others simply take the initial conditions and perturbations from the same global EPS that provides the boundary conditions – this is normally referred to as downscaling. In a downscaling EPS the forecast needs to run for a number of hours before the model can "spin up" the higher resolution detail.

3.3   Convective-scale EPS

Convective-scale NWP, with model grid lengths of 1–4 km run over relatively small domains, is now available in a number of more advanced NWP centres. These models, sometimes referred to as convection-permitting, are able to resolve some of the detail of large convective systems, and thus can attempt to predict details such as the location and intensity of thunderstorms. While this offers great potential for improved forecasts, convective systems evolve very rapidly and have short predictability timescales, so the forecasts can rapidly be affected by chaos. Ensemble Prediction Systems are therefore highly relevant to convective-scale NWP, because convective instability adds a new scale of forecast uncertainty not resolved by the lower resolution models, and with much shorter timescales.

In addition to convection itself, models on this resolution have greatly enhanced the capability for forecasting other aspects of local weather, such as low cloud and visibility of interest to aviation. Many of these phenomena are significantly affected by topographic forcing, which may give enhanced predictability when that forcing (for example, slopes, coastlines, vegetation, albedo) can be resolved by the models (for example, convective initiation or valley fog). Convective-scale EPS has the potential to provide information on the predictability of all these weather elements.

At the time of writing in 2011, convective-scale EPS are under development at various centres. In Germany, the Deutscher Wetterdienst has been running the COSMO-DE-EPS with a resolution of 2.8 km in pre-operational mode since December 2010. The United Kingdom Met Office and Météo-France have plans to introduce such systems in the near future, and research is being conducted in other countries.

Owing to the very high cost of running convective-scale EPS, they are unlikely to be available outside the producing nations for many years, and experience of them is still very limited. They are discussed only briefly in the present guidelines.

The much higher resolution of convective-scale EPS is expected to allow better resolution of many weather phenomena than is possible with global and regional EPS, for example local winds forced by topography and possibly elements such as low cloud and visibility, especially where such phenomena are forced by local details of the topography or land surface.

For precipitation, the models are likely to better resolve the intensity and spatial scales of local precipitation, especially in convective precipitation. However, to sample the full range of uncertainty in convective precipitation would require very large ensembles with hundreds or thousands of members, which will not be affordable in the foreseeable future. It is therefore strongly recommended that convective-scale EPS is post-processed using techniques such as neighbourhood processing (where it is assumed that a feature such as a convective shower may be realistic but may be misplaced and occur anywhere around the neighbourhood within, say, 10 grid lengths of where it appears in the model) to provide a more realistic spatial distribution of probabilities. Similar techniques may also be appropriate for other variables, to take account of the small size of the ensembles.

4.   STANDARD EPS PRODUCTS

This section describes some of the standard EPS products that are generated from most EPS and briefly how they may be used.

4.1   **Basic direct model output product generation**

A range of basic products is produced from most EPS directly from model output fields. These typically include the following.

4.1.1   *Ensemble mean*

This is a simple mean of the parameter value between all ensemble members. The ensemble mean normally verifies better than the control forecast by most standard verification scores (root mean squared error, mean absolute error, temporal anomaly correlation coefficient, etc.) because it smoothes out unpredictable detail and simply presents the more predictable elements of the forecast. It can provide a good guide to the element of the forecast that can be predicted with confidence, but must not be relied on its own, as it will rarely capture the risk of extreme events.

4.1.2   *Ensemble spread*

This is calculated as the (non-biased) standard deviation of a model output variable, and provides a measure of the level of uncertainty in a parameter in the forecast. It is often plotted on charts overlaid with the ensemble mean. Figure 1 shows both ensemble mean pressure at mean sea level (PMSL) as black contours and spread of PMSL as colour shading. The areas of strong colours indicate larger spread and therefore lower predictability.

4.1.3   *Basic probability*

Probability is frequently estimated as a simple proportion of the ensemble members that predict an event to occur at a particular location or grid point, for example, 2-m temperature less than 0° Celsius, or more than one standard deviation below normal. Figure 2 shows the contoured probability of wind gusts exceeding 40 kt. The ensemble mean PMSL is also included as grey contours.
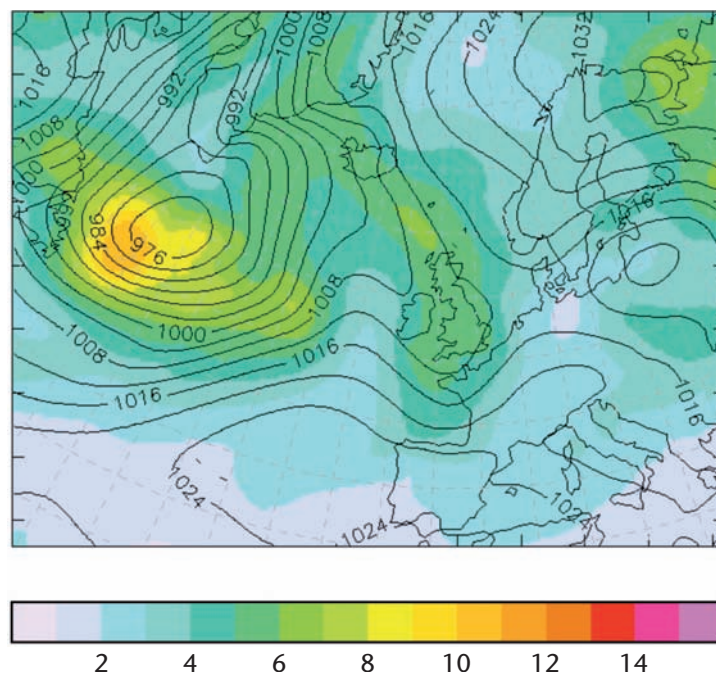
Source: UK Met Office, © British Crown Copyright

Figure 1. Mean (black contours) and spread (colour shading) for PMSL forecast (T + 72)

No members   0.01 0.1 0.2   0.4   0.6   0.8 0.9 0.99 All members
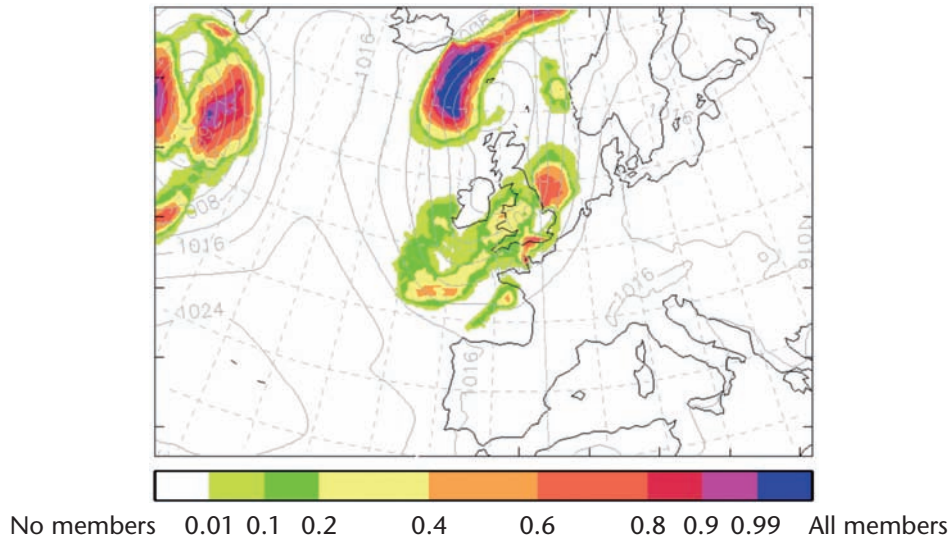
Figure 2. Regional MOGREPS probability map for gust speed > 40 kt for 16 July 2010 at 0300 UTC (T + 21 from 15 July 2010 at 0600 UTC); ensemble mean PMSL plotted as faint background

It should be noted that this definition of probability is not a true Bayesian probability as would be defined by a statistician, but provides a useful estimate for practical purposes. It makes an assumption that the model accurately reflects the climate distribution of occurrence of an event. Probability forecasts produced in this way should always be verified over large samples of cases to determine the extent to which forecast probabilities relate to observed frequencies.

The example in Figure 3 was produced for the project in the South Pacific, part of the larger SWFDP.

0.01   0.25   0.5   0.75   0.99
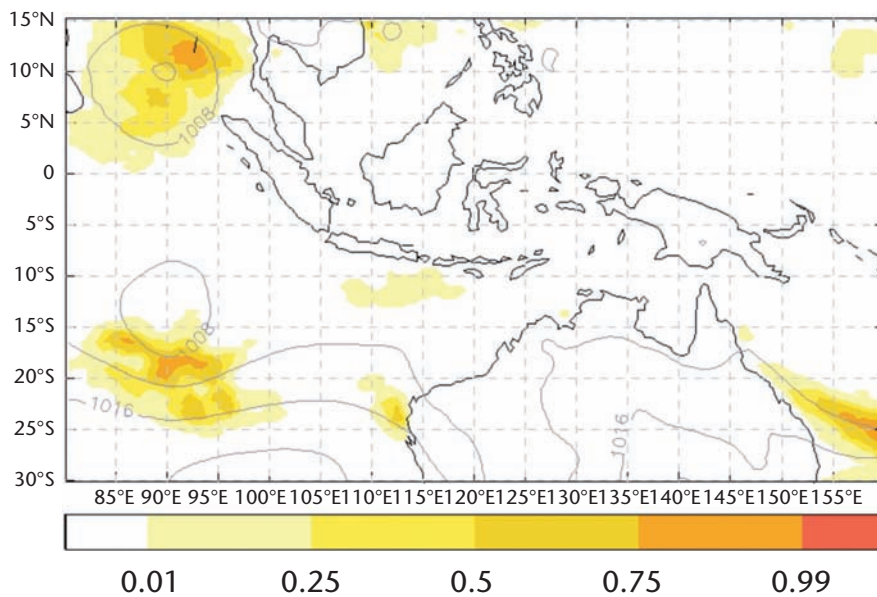
Figure 3. Global MOGREPS probability map for 10-m wind speed > 20 kt for 5 November 2010 at 0000 UTC (T + 48 from 3 November 2010 at 0000 UTC); ensemble mean PMSL plotted as faint background

4.1.4   *Quantiles*

A set of quantiles of the ensemble distribution can provide a short summary of the uncertainty. Commonly used quantiles are the maximum and minimum of the ensemble distribution, and the 25th, 50th (median) and 75th percentiles. Others often used include the 5th, 10th, 90th and 95th percentiles.

4.1.5   *Spaghetti maps*

Charts showing a few selected contours of variables (for example, 528, 546 and 564 Dm contours of 500 hPa geopotential height) from all ensemble members can provide a useful image of the predictability of the field. Where all ensemble member contours lie close together the predictability is higher; where they look like spaghetti on a plate, there is less predictability (see Figure 4).

4.1.6   *Postage stamp maps*

A set of small maps showing contoured plots of each ensemble individual member (see Figure 5) allows the forecaster to view the scenarios in each member forecast and assess the possible risks of extreme events. However, this presents a large amount of information that can be difficult to assimilate.

4.1.7   *Site-specific meteograms*

Model output variables can be extracted from the grid for specific locations. There are many presentations that can be used to represent the forecast at locations, such as plume charts and probability of precipitation. One of the most commonly used is the ensemble meteogram (or EPSgram) which uses a box and whisker plot to illustrate the main percentile points of the forecast distribution for one or more variables (see Figure 6).
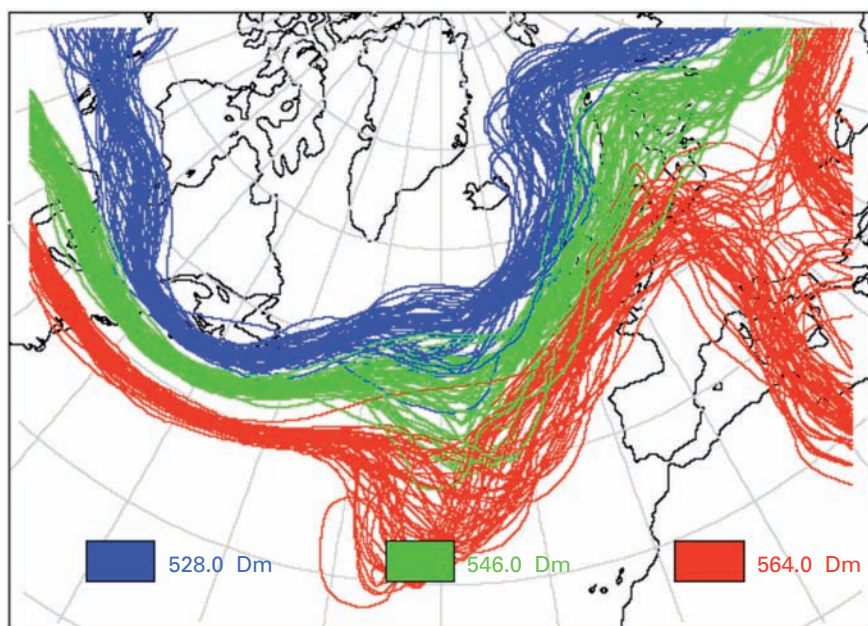
Source: UK Met Office using data from ECMWF, © British Crown Copyright

Figure 4. Ensemble 500 hPa forecast spaghetti charts for 11 February 2001 at 1200 UTC (T + 96 from 7 February 2001 at 1200 UTC)
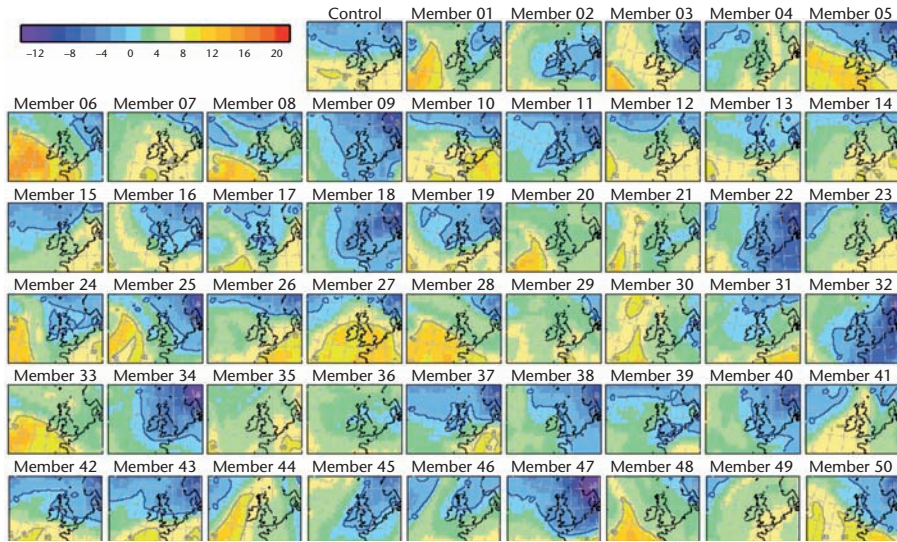
Source: UK Met Office using data from ECMWF, © British Crown Copyright

Figure 5. Postage stamp map for 7 February 2009 at 1200 UTC (850 hPa wetbulb potential temperature, in degrees Celsius; T + 300 from 26 January at 0000 UTC)

5. GENERAL COMMENTS APPLYING TO ALL USE OF EPS

This section presents a number of general principles that apply to all uses of EPS. Those sections that follow provide more detail on the specific use of EPS for particular types of forecast production.
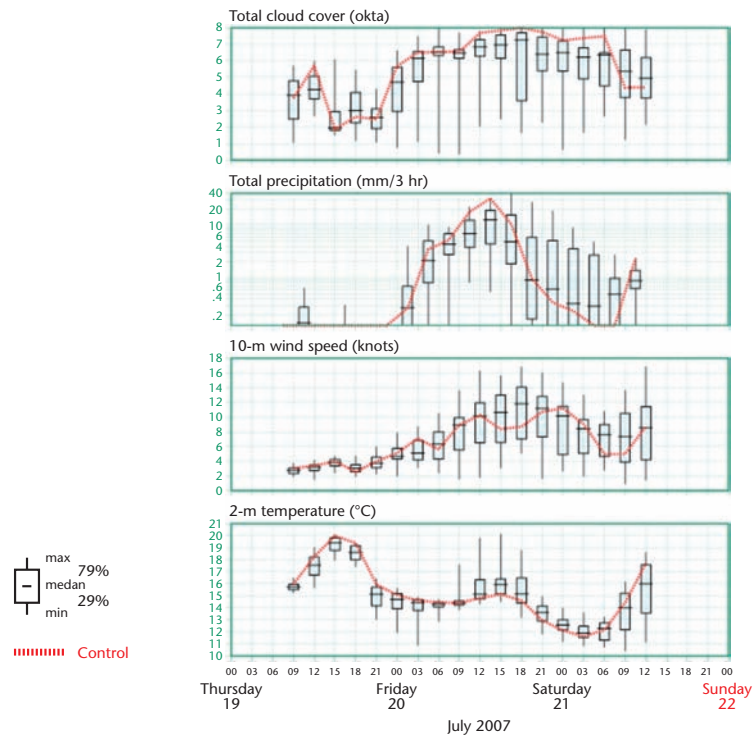
Source: UK Met Office, © British Crown Copyright

Figure 6. MOGREPS European EPS Meteogram for Brize Norton (51.8°N 1.6°W) from 19 July 2007 at 0900 UTC to 21 July 2007 at 1200 UTC

(a)   An EPS best represents the uncertainty in resolved variables.

    (i)   Upper-air usually more skilful than surface

        –   Surface parameters are affected by sub-grid scale uncertainty not resolved by the model.

    (ii)   As resolution and model performance increases, the ability to predict surface weather parameters is continually improving.

(b)   An EPS is only as good as the model(s) it uses.

    (i)   If a model is unable to represent certain phenomena, the EPS will also be unable to represent it.

        –   A good example is that most ensembles cannot resolve convective storms, which is one of the reasons why some centres are developing ensembles at convective scale.

    (ii)   An EPS will share any systematic biases of the model used.

(c)   How to combine deterministic forecast with ensemble/probabilistic?

    (i)   Relative capabilities of ensemble members compared to high resolution/control.

    (ii)   See the Guidelines on using information from EPS in combination with single higher resolution NWP forecasts (February 2006).

(d)   A common question is whether a forecaster can improve the distribution by re-weighting members (for example, the high-resolution control forecast if included) or by rejecting some members.

    (i)   Forecasters may think that some members are unrealistic.

    (ii)   Can we eliminate some members on the basis of recent observations or pick a "best member"?

        –   PERHAPS, for certain aspects of the forecast over very short-period forecasts and for local forecasts over a small area.
        –   Over a large area or the full model domain, the control forecast will always be the most skilful.
        –   NOT for longer-period forecasts.

    (iii)   This type of approach is subjective and difficult.

    (iv)   It is strongly recommended that forecasters use the whole EPS distribution in a probabilistic approach.

(e)   Strengths and weaknesses of the models/ensembles available to the forecaster should be known. Documentation should be easily available to the forecaster.

    (i)   Verification of multiple thresholds to be available

    (ii)   Summary documentation of strengths and weaknesses by season.

(f)   Be careful with "end-of-chain" diagnostic parameters, for example, precipitation and cloudiness. For instance, look at distributions of indices in convective situations.

(g) Forecasters should not always rely on direct model output of weather variables, but should also consider analysis of better resolved diagnostics which may aid interpretation of the EPS forecast (for example, synoptic features, environment/precursors/potential for high-impact weather developments such as moisture convergence, low-level jets, development regions and convective diagnostics).

The use of EPS (and other probabilistic tools) opens the possibility of issuing two different types of forecasts, fully probabilistic, or deterministic with supplementary uncertainty information (for instance, confidence). Which type is used affects who makes decisions from the forecast. In general, the use of fully probabilistic forecasts allows all users to tailor their decision to their specific needs (for example, using cost-loss estimation), and is therefore strongly encouraged.

6. USE OF EPS IN DETERMINISTIC FORECASTING

In general, it is strongly recommended that probabilistic forecasts provide the best and most complete weather forecast for customers, and should be encouraged, especially at longer lead times. However, it is recognized that many customers demand a simple deterministic forecast, and where a deterministic forecast is to be produced, the use of an EPS can often provide a more reliable forecast than a single deterministic NWP run. This is particularly true for forecasts more than 1–3 days ahead, and can help reduce jumpiness from run-to-run of the forecast system at any time range.

Several indicators from the EPS can be used to optimize the deterministic forecast. The ensemble mean will on average score the best by many standard verification scores, but it must be remembered that it will tend to smooth out the smaller scale unpredictable detail, and will rarely capture the intensity of important high-impact weather systems. Thus, the ensemble mean should not be used on its own if the forecast will be used for potential severe weather impacts. Other useful guides to the most likely forecast can be the median (central point in the pdf) or mode (most likely value in the pdf) – these are easier to identify for single weather parameters than for the complete forecast picture.

If a deterministic forecast is to be issued, it may sometimes be augmented by a statement of the confidence of this forecast to take some advantage of the uncertainty information available. The confidence will not always be the same for all elements of the same forecast. Confidence indices, if used, are best provided separately for each variable. The confidence level should be based on the spread of the ensemble, but also considering the known forecast skill limitations.

The best approach to issuing a deterministic forecast will depend on the predictability as indicated by the ensemble spread. The spread could be analysed using various products such as spaghetti plots and a map depicting variance at the synoptic scale and then, at the lower scales, using meteograms, quantiles, cluster analysis, and so on.

(a) Small spread in the ensemble (good predictability).

    (i) In this case it may be reasonable to offer more detail in the forecast.

    (ii) Take the control, the high-resolution control, the ensemble mean or the median as a guide (with due regard for the need for calibration or bias correction).

    (iii) Spread may often differ between model variables so small spread in one parameter does not guarantee confidence in all aspects of the forecast.

        – Good synoptic scale predictability does not always mean predictability in surface weather variables such as temperature or convective precipitation.
        – Forecaster should still take account of uncertainty in parameters not resolved by the model.

(b)   Large spread in the ensemble (poor predictability).

    (i)   Avoid giving too much detail in the forecast.

    (ii)   Ensemble mean should be considered, but if the ensemble covers a range of scenarios the ensemble mean will not provide a realistic scenario.

    (iii)   In that situation, take the most representative member of the ensemble (for example, most populated cluster or mode of pdf) as a guide to the most probable outcome.

        –   Note that the most representative ensemble member may not give the most probable value for each weather element (for example, most probable temperature at a location may not be correlated with the most probable precipitation amount).

    (iv)   The uncertainty assessment.

        –   Encourage users to follow forecast updates.

    (v)   Take into account extremes of the EPS and of the high-resolution control.

        –   Make a careful evaluation of the possible evolutions of the synoptic situation and their potential impacts.
        –   Take into account the behaviour of models.
        –   The high-resolution control may be better able to represent certain high-impact events.

(c)   In the short range (12–18 hours), it may be possible to take into account the latest observations (3–6 hours into the forecast) in order to choose a scenario or a member of the ensemble.

    (i)   For example, a rapidly evolving cyclone may be best predicted by the member with the best position after a few hours but ONLY in the very short range.

    (ii)   Be aware that future evolution will be influenced by features coming from upstream. This makes member selection for forecasts beyond ~24 hours impossible.

    (iii)   Also the consistency of the latest runs with respect to the previous is a factor to take into account.

(d)   In the longer range, while probabilistic forecasts are best suited, if a deterministic forecast is to be produced, the use of the ensemble mean or median could yield more reliable forecasts, with less jumpiness between runs of the forecast.

6.1   **Decision-making from deterministic forecasts**

Weather forecasts are only useful when they are used to make decisions. It is often argued that it is easier to make a decision from a deterministic forecast than a probabilistic one. However, when the forecaster issues a deterministic forecast the underlying uncertainty is still there, and the forecaster has to make a best guess at the likely outcome. Unless the forecaster fully understands the decision that the user is going to make based on the forecast, and the impact of different outcomes, then the forecaster's best guess may not be well tuned to the real needs of the user.

(a)   Some knowledge of the needs of the end-user is required when choosing to make a deterministic forecast for a specific event to occur. An optimal decision cannot be made without the cost–lost ratio of the user. This ratio can be assessed by a survey or a direct discussion with the end-user.

(b)   When appropriate, the forecasters should convey the risks and impacts associated with worst-case scenarios alongside the most likely outcome.

7.   **SCENARIOS**

A useful way to summarize the uncertainty in a weather forecast can be to describe a small number of possible outcomes or scenarios rather than giving the full detail of a probabilistic forecast. For some customers used to receiving deterministic forecasts, this may be more acceptable. Ideally the EPS can be used to estimate the relative likelihood of the different scenarios presented. In most cases, to avoid confusion, the best approach may be to issue a most-likely scenario based on the advice above on issuing deterministic forecasts, plus a single alternative scenario. This may often be a worst-case scenario, perhaps reflecting a low probability but high-impact possibility suggested by the most extreme ensemble members. However, care should be taken not to give the impression that either scenario will be correct, as the truth could easily lie somewhere in between (or even be different again!).

Useful tools to aid in providing alternative scenarios are postage-stamp maps (see 4.1.6 above), which show the forecaster all the individual forecasts in the ensemble, or clustering (see 9.3 below), which automatically groups the ensemble members and provides the forecaster with an objective assessment of the possible scenarios.

8.   **FULL PROBABILISTIC FORECASTS**

Wherever possible, the use of a full probabilistic approach is recommended in issuing forecasts. This provides a full representation of the uncertainty information provided by an EPS, and also allows users to tune their decision-making to take account of their particular applications.

Probabilistic forecasts can be expressed in a number of ways, and need not always use the word probability, such as:

(a)   A forecast of a weather variable provided with error bars that vary according to the ensemble spread;

(b)   A fuller representation of the ensemble distribution showing a number of percentile values, as used in the standard meteogram product;

(c)   Probabilities of specific (well-defined) events occurring, expressed as numbers or as contoured shading on a map.

When a forecast is presented as a probability, it is very important to express very clearly what the probability is for, so that it is clear and understandable to both the forecaster and the user. We often talk about the probability of an event occurring, and it is this event that must be defined. Often the event will be for a threshold value to be exceeded (for example, more than 50 mm of rain, or temperature below 0° Celsius). Ideally it will be something that has an important impact for which someone will have to take a decision (for example, the probability that ice will form on roads so that road treatment will be required). It is also important to define when and where the event is forecast for:

(a)   Exact time, or time period to which the forecast refers.

(b)   Exact location or area to which the forecast applies.

       –   If it is an area, does it mean a forecast that the threshold will be exceeded somewhere in the area, or everywhere in the area?

A good test of whether an event is well defined is to ask yourself whether you could easily measure whether the event does happen or not (in other words, could you verify the forecast). If you cannot easily say, then you may need to define the event better.

The following provides a number of issues that should be considered when basing probabilistic forecasts on EPS outputs:

(a)   Calibrated, bias corrected forecast can be directly issued to the end-user (low cost).

–   This approach allows for the possibility of issuing automated forecasts for many locations and users.

–   Methods for bias correction and calibration are discussed in section 9.

(b)   Direct model output from ensembles should be used with care, as it may not provide reliable probabilistic forecasts, but nevertheless will often provide valuable information. In some cases use of DMO may be essential where there is no calibration system in place – calibration is difficult for certain variables such as precipitation, or where adequate observations are not available.

(c)   To generate probabilistic forecasts of outcomes dependent on more than one weather element, it is important to calculate this outcome for each ensemble member and then combine members to create the probabilities. This retains consistent correlations between different weather variables and also different locations (for example, the correlation in temperature between two locations). Calibration or post-processing may spoil this consistency.

–   This principal also applies when using the ensemble to drive downstream impact models (for example, hydrological models) where the downstream model should be run for each ensemble member and then the probability of the downstream impact calculated.

(d)   In "usual" situations, forecasters should not try to change the probabilistic forecasts issued by the EPS (DMO or post-processed). The forecasts can be issued directly to the public. Forecasters should target their attention to "unusual" situations.

(e)   In "unusual" situations, probabilistic forecasts can be adapted by the forecasters using, for example, experience, analogues and conceptual models. Forecasters may be able to correct for some known system biases or model weaknesses. The corrections should be made by using the guidelines mentioned in section 9.

(f)   Studies have shown that the general public is able to make better decisions when presented with uncertainty information in forecasts than with a deterministic forecast. When uncertainty information is not provided people make their own assumptions.

(g)   Probabilities have to be presented in a comprehensive graphical form. For examples and guidelines, see the *Guidelines on Communicating Forecast Uncertainty* (PWS-18, WMO/TD-No. 1422).

(h)   Probabilities of events relevant to specific applications should be defined. This includes, for example, application in agriculture in which the occurrence of dry spells or rainy periods influences irrigation, seeding and harvest.

(i)   Risk is a combination of the likelihood of a phenomenon and its potential impact that can be estimated by the EPS. It gives an objective and valuable decision basis to the forecasters in order to assess different warning levels. Impact has to be agreed with the relevant authorities (Public Weather Services customers). Climatology usually provides a good reference to establish the thresholds of phenomena that produce impact. The thresholds can

be adapted taking into account the recent evolution of the various environmental parameters (recent rainfall accumulations affect soil saturation, leaf-cover on vegetation, snow cover, etc.).

(j)   It is recommended that where probabilities are indicated for significant high-impact weather, the forecaster should add a written comment or warning.

9.   POST-PROCESSING

The aim of the present guidelines is to provide explanation and advice for post-processing using statistical dynamical and other approaches to improve EPS outputs. There are numerous approaches and some of the most common are described in this section. Some methods are quite generic and may be best applied by EPS producers at source, while others are quite specific to applications and may be better applied specifically for individual users.

9.1   Statistical post-processing

Generally speaking, statistical post-processing is needed in order to correct systematic errors in models and thereby add value to direct NWP model output. These errors are particularly important for surface parameters (for example, 2-m temperature, 2-m humidity, 10-m wind speed, precipitation and total cloudiness) and are linked to local conditions.

More precisely, statistical post-processing can be used to:

(a)   Remove systematic biases;

(b)   Adjust ensemble spread;

(c)   Quantify uncertainty not represented directly by the EPS;

(d)   Predict what the model does not represent explicitly (for example, low visibility).

In general statistical methods are easier to apply to some types of model output variables than others. Temperature is often relatively easy, for example, as it is a continuous variable and varies relatively smoothly in model fields, and most importantly temperature errors are often approximately normally distributed. Precipitation, by contrast, is particularly difficult because precipitation fields often have much multi-scale structure that is poorly represented by models, especially on the small scales. Its climatological distribution, and hence the distribution of forecast errors, is bounded at zero at one end and often highly skewed, making it much more difficult to represent statistically. The problem can sometimes be reduced by transforming the distribution to make it more quasi-normal, but in general post-processing methods for precipitation are much less effective than for other variables.

9.1.1   *Bias correction of the first moment of the probability distribution function*

This post-processing is similar to Model Output Statistics (MOS) methods applied for single models, but with some important differences. For ensembles, it is well known that a traditional MOS which is trained specifically for each forecast lead time will lead to a significant decrease of the ensemble spread at longer lead times. Instead, it is recommended to use a pseudo-perfect prognosis approach. This method is based on the use of MOS statistical models computed over the first 24 hours of the forecast and then applied to the corresponding steps at all forecast lead times.

Adaptive methods such as the Kalman filter are recommended to allow the corrections to be automatically updated to account for model changes (upgrades) and changes in the season.

In the case of single-model ensembles (that is, the same model is used for all of the members, even where model perturbations are implemented) the same statistical model should be trained using the control forecast and applied to all members of the ensemble.

In the case of multi-model or multi-physics ensembles (that is, where different models are used to build the pdf, or systematically different model versions are applied, for example, different parameterization schemes) specific statistical models should be trained and applied for each model version.

In either case, the development of these statistical models needs a training set of model outputs (predictors) and observations (predictands). In the case of adaptive methods such as the Kalman filter, this training set is updated continuously from the daily forecasts.

The "observations" can be either site-specific observations or may be the best available set of analyses. In the case of site observations the statistical post-processing will lead to local forecasts (that is, at each site-specific point where observations are available). When analyses are used the end product is a bias-corrected and downscaled gridded forecast.

It should be noted that when different weather variables are independently bias-corrected, some of the correlation between variables represented by the different ensemble members may be lost. For this reason forecasters may prefer to view direct model outputs.

9.1.2   *Calibration of higher moments of the probability distribution function*

Bias removal for the second moment of the pdf is often known as "calibration". It aims to improve the reliability of the probabilistic forecast. Therefore this kind of post-processing is specific to EPS and is particularly important to optimize probability forecasts. As for the first moment bias correction, calibration is based on local conditions and requires high-quality observations or analyses as a reference.

A number of methods are under development that attempt to calibrate both the first and second moments of the pdf to optimize the complete distribution, including the following:

(a)   A method developed at the University of Washington in the United States of America is now considered as one of the best to deal with this issue. This method, called Bayesian model averaging, is based on specific statistical assumptions, for example, normal distribution for temperature;

(b)   The Ensemble Kernel Distribution Model Output Statistics (EKDMOS) is another technique that has been implemented.

The above methods are commonly applied to variables such as temperature and wind speed. Variables such as precipitation are more difficult to correct because of the nature of the pdf and the local variability of observations. Some specific approaches are under development, but post-processing methods are at present less successful and may not improve significantly over raw model outputs.

It must be noted that there are limitations to the potential of statistical post-processing especially in the case of severe events. Commonly calibration will improve the statistical reliability of probabilistic forecasts (the match of forecast probabilities to frequency of observations of the event) but reduce the resolution of the forecasts (the ability to discriminate whether an event will occur or not). Sometimes it is found that calibration will improve forecasts of common events, but degrade the probabilities of more extreme events. The main reason for this is that observations of these kinds of events are rare, and the statistical distributions are trained to the more common

events. Therefore calibration cannot be expected to provide significant improvement over the raw forecasts in this case.

Some attempts have been made to develop post-processing explicitly for prediction of more extreme events, for example first-guess severe weather warning systems. In these cases the systems can be calibrated specifically to optimize the reliability for extreme thresholds. Nevertheless, human expert interpretation remains particularly important for assessment of the risk of extreme events.

9.2   **Downscaling**

A number of methods may be used to add some local detail to forecasts generated with lower resolution models, and these techniques may be applied to EPS forecasts just as with deterministic NWP.

9.2.1   *Dynamical downscaling*

Dynamical downscaling may be defined as the use of a higher resolution limited-area NWP model to add detail forced by topographic detail and to resolve fine-scale processes such as convection. Ideally all ensemble members will be downscaled, but where cost constraints prevent this, a selected set of members may be downscaled. In dynamical downscaling, the initial conditions, boundary conditions and perturbations are taken directly from the lower resolution EPS members. Care must be taken to ensure that the downscaling is appropriate to ensure good performance of the high-resolution model, for example, appropriate ratios of grid sizes and rate of updating of boundary conditions. The model performance should be carefully tested over the domain. Many LAM and convective-scale EPS are dynamical downscaling systems from global ensembles.

9.2.2   *Topographic downscaling using simple physical models*

For some parameters such as 2-m temperature and 10-m wind speed a simple downscaling can be applied using a relationship to the surface topography. For example, in surface temperature forecasts the lapse rate may be used to downscale the low-resolution EPS field to a higher resolution grid using a gridded topography. Figure 7 shows probabilities of strong wind downscaled from a regional EPS using a high-resolution orography field, and shows how probabilities of winds over the mountains in Scotland can be detected which were missed in the DMO version of the chart.

9.2.3   *Site-specific extractions*

Forecasts for specific locations may be generated by extracting data from model grids. In the simplest implementations data are simply taken from the nearest model grid point, or are interpolated between the nearest grid points by linear interpolation. Various methods are used to improve on these approaches, using similar techniques to the downscaling methods. In particular, corrections to surface temperature and wind speed should be made to account for the difference between model orography and the true altitude of the site. An intelligent grid-point selection system which chooses the most representative grid point can also be better than a simple interpolation, especially near coastlines where it may be better to choose the nearest land point to represent a land location, rather than for example the nearest grid point, which may be over the sea. This approach may also be beneficial near steep orography.

A one-dimensional model could also be used for specific forecast applications, for example, 1D fog models for airports.
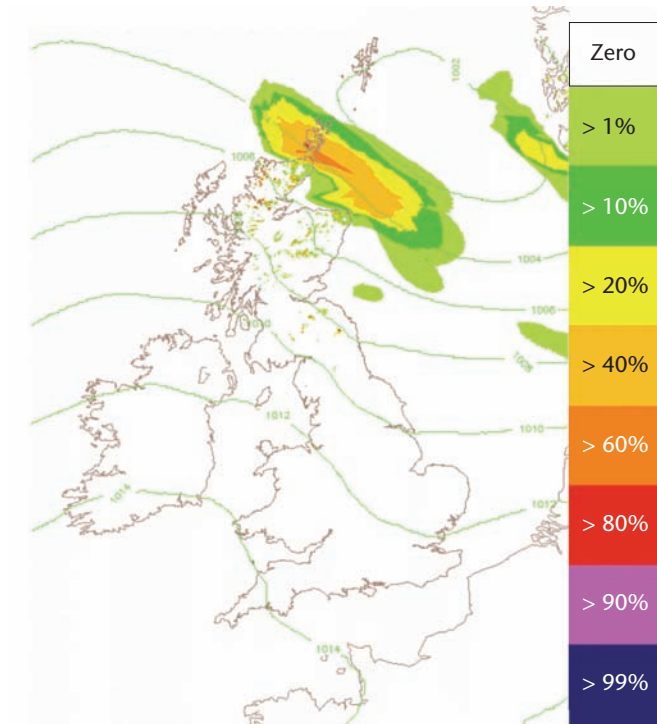
Source: UK Met Office, © British Crown Copyright

Figure 7. Probabilities of strong wind for 5 August 2011 at 0900 UTC (T + 15) calculated from MOGREPS-R after downscaling model output fields to a 2-km grid using a high-resolution orography field. Note the high-resolution probabilities of strong wind over the mountains of NE Scotland revealed by the downscaling.

9.2.4   *Statistical downscaling*

Downscaling of surface fields may also be done by building a statistical relationship between low-resolution model fields and high-resolution analyses. Two approaches that may be followed are given below.

9.2.4.1   Using analysis differences

The statistical relationship may be developed by comparing high-resolution gridded analyses with the corresponding analysis fields on the EPS model grid. This provides a downscaling vector, which may then be applied to EPS forecast fields to provide bias-corrected and downscaled forecast fields on the high-resolution grid.

9.2.4.2   Kalman filter

A Kalman filter approach may be applied at each grid point of the high-resolution grid to build a statistical relationship with the lower resolution EPS analysis fields. This Kalman filter may then be applied to the EPS forecast fields to provide bias-corrected and downscaled forecast fields on the high-resolution grid.

9.2.5   *High-impact weather diagnostics*

A number of methods are available to diagnose specific high-impact weather phenomena from NWP models, and these can be applied equally to EPS. A good example is severe convection

diagnostics. These often use multi-level outputs of several models to diagnose the instability and potential for severe convection, and provide probabilities for phenomena such as large hail, tornadoes and convective wind gusts.

9.2.6 *Downscaling by combination of low-resolution EPS and high-resolution control forecast*

Low-resolution ensemble perturbation fields (difference between the perturbed member forecast and the control forecast) can be added to high-resolution control forecast fields to provide a high-resolution probabilistic forecast.

9.3 Clustering techniques

Classification processes can be used to synthesize the vast amount of information contained in ensembles. Different kinds of classifications can be implemented, as follows:

(a)   Clustering attempts to group together members that are most similar in their evolution over a defined geographical region of interest. Several standard clustering algorithms are available and may produce different results. The clustering outcome also depends on the variables chosen;

(b)   The "tubing" classification identifies a central cluster of the members closest to the ensemble mean and those members most significantly different from the ensemble mean (tube extremes). Tubing is useful to identify the most likely outcome and also the possible scenarios most different from that solution;

(c)   Classification of forecasts by matching ensemble members to a defined set of flow regimes, for example, the Grosswetterlagen types defined for central Europe. This method may provide the clustering which best matches a synoptic forecaster's expectations.

9.4 Use of reforecasts

Research has shown that calibration of ensemble forecasts using historical sets of reforecasts – forecasts run with the same model or EPS from sets of historical cases, initiated from reanalyses – can be very effective in improving the quality and reliability of probabilistic forecasts. Such reforecasts provide a better dataset for training of statistical post-processing methods compared to using recent forecasts, as they provide a better sampling of different weather regimes and types. This can be particularly useful for optimizing the calibration of forecasts for rare or extreme events. However, the running of reforecasts adds substantially to the computing cost of running an EPS, and depends also on the availability of a suitable reanalysis dataset to provide the initial conditions. As a result very few EPS currently have reforecast datasets available, but their use is recommended where possible. Where a full reforecast dataset is not available, an alternative may be to use a recent archive of EPS forecasts from the same system, although this is likely to provide a less reliable sampling of the full model climate.

9.4.1 *Extreme Forecast Index*

One application of reforecasts is the computation of an Extreme Forecast Index (EFI).

NWP models and EPS do not represent accurately the climate of the real atmosphere, and identification of extreme events may be best done in relation to model climatology. The Extreme Forecast Index developed by ECMWF (see Figure 8) allows identification of forecasts that are extreme relative to the model climate, providing an alert to a risk of severe weather, but does not provide explicit probabilities of severe events.
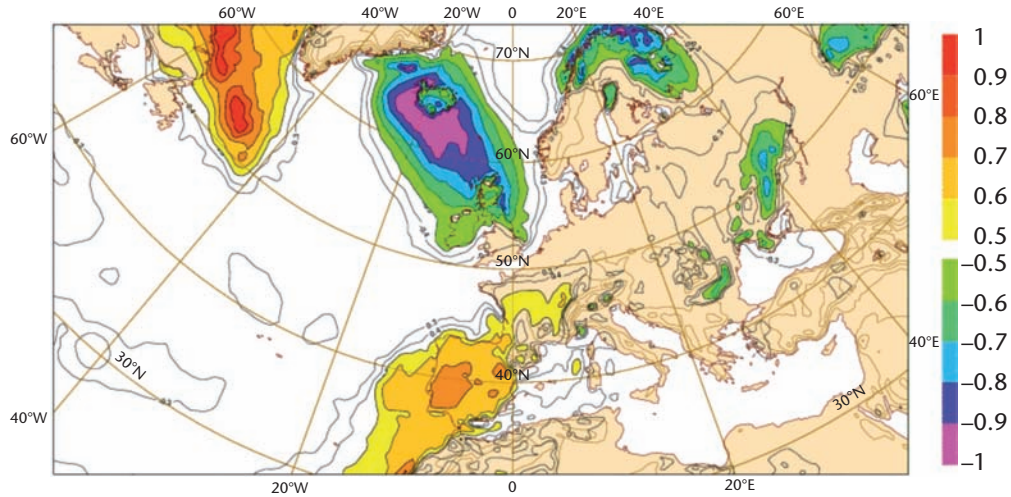
Source: ECMWF

Figure 8. Extreme forecast index (2-m temperature) for the period 6–7 January 2011 at 0000 to 0000 UTC (T + 60–84 from 3 January 2011 at 1200 UTC)

Reforecasts can also be used to assess forecast severity in relation to climatological return periods, which can be a useful way to communicate the severity of an event.

9.4.2   ***Quantile–quantile matching***

Another approach to forecast calibration which can be used where an estimate of the model climate is available is quantile matching. For example, the value corresponding to the 90th percentile of the model climate may be interpreted to represent the 90th percentile of the real observed climate distribution for a particular location. In general this method requires the use of a reforecast dataset to provide the model climate.

9.5   **Feature tracking**

A useful technique for lower resolution EPS such as global EPS is to track meteorological features in each member of the ensemble. Although not well resolved in the model, tropical cyclones are a good example of a meteorological feature whose movement is nevertheless well predicted by global models. A global EPS could not be expected to predict the intensity of strong winds or heavy rain in a tropical cyclone but could track its position. The forecaster can interpret the probabilities of severe weather by knowing the characteristics of tropical cyclones, combined with the ensemble information on where it is likely to go. Figure 9 shows tracks of Hurricane *Tomas* in the members of the ensemble (left), probabilities that the storm will pass close to locations on the map (centre), and summary tracks such as the ensemble mean track (right). These types of charts are often made available to the tropical cyclone Regional Specialized Meteorological Centres.

10.   **USE OF EPS IN PREDICTION OF SEVERE WEATHER AND ISSUE OF WARNINGS**

Severe or high-impact weather events occur on a wide range of scales in space and time, from tropical cyclone, extra-tropical cyclone, monsoon, winter storms and other large-scale systems, to smaller scale systems such as local severe storms, orographic precipitation, thunderstorms and tornados. Forecasters must take account of the different predictabilities of different types of events (for example, do not try to predict a thunderstorm three days in advance).

A well-structured National Meteorological and Hydrological Service (NMHS) severe weather warning system should have appropriate thresholds, lead times and level of service agreed with users. Thresholds should normally reflect the level of impact the weather is expected to have on society, including danger to life and property, and disruption to everyday life. Features which should be considered in a warning system include the following:

(a)   Types of warnings; regions; thresholds (severity/impact and probability).

   (i)   Risk = probability x impact

(b)   A good warning system is one that will be easily understood by users, with standard thresholds adhered to by forecasters.

   (i)   Many countries now use a four-colour traffic light system (green, yellow, amber and red) indicating different levels of risk and corresponding levels of action that users should take.

(c)   A good warning system will require feedback from users to NMHSs. In turn, NMHSs should give feedback to producers enabling them to design appropriate products.

Ensemble Prediction Systems are a powerful tool in predicting severe weather events. For impact-based warnings systems the EPS may be used to help estimate the probability of weather hazards for use in the estimate of risk = probability x impact. Nevertheless, EPS can specifically predict severe weather for which the model or models have been able to resolve. Otherwise, the following applies:

(a)   Numerical weather prediction has limitations in explicitly resolving smaller scale phenomena, which leads to underestimation of extreme events likelihood within EPS;

(b)   EPS sometimes can identify precursor conditions for severe developments or favourable large-scale environment such as convective indices;

(c)   Lower resolution EPS (global) are less likely to be able to resolve details of an extreme event;

(d)   Regional EPS, which usually have higher resolution, should provide more detailed uncertainty estimates at the smaller scales.
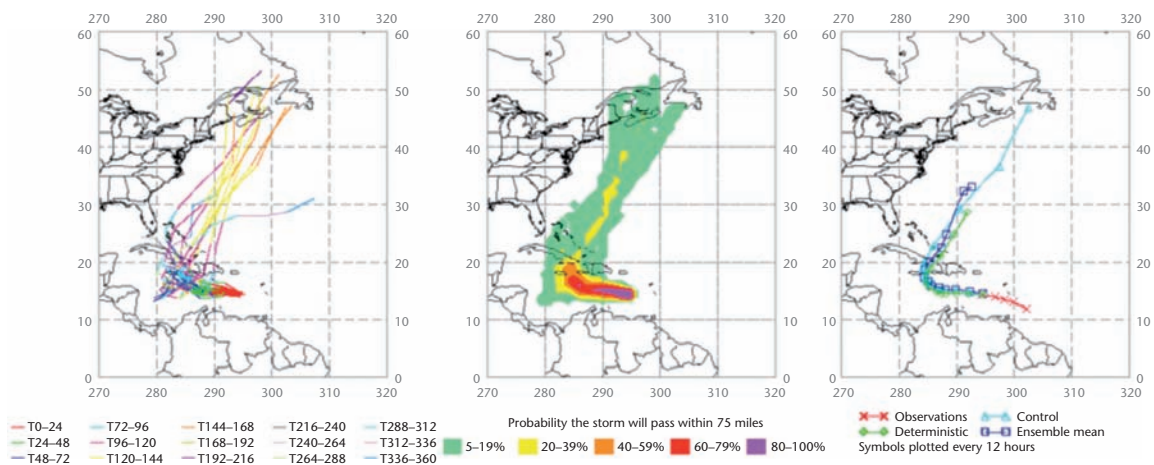
Source: UK Met Office, © British Crown Copyright

Figure 9. Tropical cyclone products from the MOGREPS 15-day ensemble showing tracks of Hurricane *Tomas* from forecast issued on 1 November 2010

Hazard thresholds used in the EPS may need to be calibrated to take account of the above limitations.

Early indications of some extreme events will be predicted in the tail of the ensemble distribution.

(a)   Therefore forecasters and users should not ignore low probability events, especially when those events are very rare.

    (i)   For example, ignoring probabilities below 20 per cent or even 10 per cent could result in missing the most important events signaled by the EPS.

    (ii)   To be able to use low probabilities, forecasters need verification information.

    (iii)   "False alarms" are actually correct features of low probabilities. However, low probabilities may be required in potential high-impact situations.

    (iv)   It is expected that the probability will increase closer to the event – usually but not always.

An extreme event may also be forecast essentially correctly, but with errors or uncertainties in location or timing.

Synoptic interpretation (for example, weather feature tracking, use of analogues) or statistical downscaling tools are ways to add skill to the basic EPS.

(a)   Note that some statistical methods require large data samples for training, and may not be well suited to rare or extreme events.

(b)   Cyclone tracking products (for both tropical and extra-tropical cyclones) can provide a useful summary of the development of high-impact storms.

(c)   There is potential for development of more feature-based diagnostics for poorly resolved severe weather systems.

The Extreme Forecast Index can be a useful tool in alerting forecasters to a potential severe event.

(a)   EFI does not provide explicit probabilities of specific events, and should be interpreted in conjunction with other tools.

(b)   Currently only a small number of systems can provide an EFI because of the need for a model climatology.

Consideration of input from multiple forecasting systems (EPS and deterministic) may give additional information on the probability of extreme events.

(a)   Production of verification highlighting the skill and limitations of EPS is important.

    (i)   Users of EPS should be aware of those limitations and strengths.

    (ii)   However, because of the rarity of most extreme events it is often impossible to provide reliable (or statistically valid) verification of probabilistic performance. It may be possible to gain some estimate of skill for extreme events by extrapolating from the verification of less-severe events.

(b)   Given the diminishing of the EPS skill with increasing lead time, the latest available products are generally given higher credibility. However, previous runs of an EPS may still provide useful information about a rare event because of the lack of spread (limitation in the sample size).

11.   SEVERE WEATHER IMPACT MODELLING

The uncertainty in the weather forecast can be propagated through to uncertainty in impact by coupling ensemble members to impact models and generating a distribution of impact predictions. Examples include hydrological models for probabilistic flood forecasting, coastal storm surge models and heat health models. This is an advanced application, which is being increasingly applied in the more advanced centres. Figure 10 shows an ensemble forecast of storm surge at a coastal port, where the weather forecasting EPS has been used to force an ensemble with a storm surge model. The red lines at the top of the graph show the flood danger level oscillating up and down with the tide, and a flood risk is indicated where the ensemble forecast surge lines cross above the red lines. This is an interesting example as one member of the ensemble produces an extreme surge at day 7, indicating a low probability of severe coastal flooding. In this situation the user needs to be able to take some early preparedness action but without overreacting because the probability of the flooding occurring is low.

12.   VERIFICATION

Verification is a very important part of everything we do in forecasting. If we do not verify our forecasts – measure how good they are by looking back afterwards and seeing how well the forecast matched what actually happened – then we have no way of learning and improving our forecasts in the future. This is just as true with probabilistic forecasts. You will often find people say that a probability forecast can never be wrong (unless we say 0 per cent or 100 per cent). Some people will also say that it is just a way for the forecaster to avoid making a decision. The way to challenge these views is to demonstrate that we do verify the forecasts, and that they have useful skill.

A detailed guide on verification of forecasts is not provided here, but a few important points are described:

(a)   A single probability forecast cannot be right or wrong.

   (i)   If we predict something with a high probability and it happens, it is often tempting to say "Look, we got it right!" We should avoid doing this, because when we forecast something with a low probability and it happens we will want to say to the user "We did say it was a possibility even though it was a low probability".
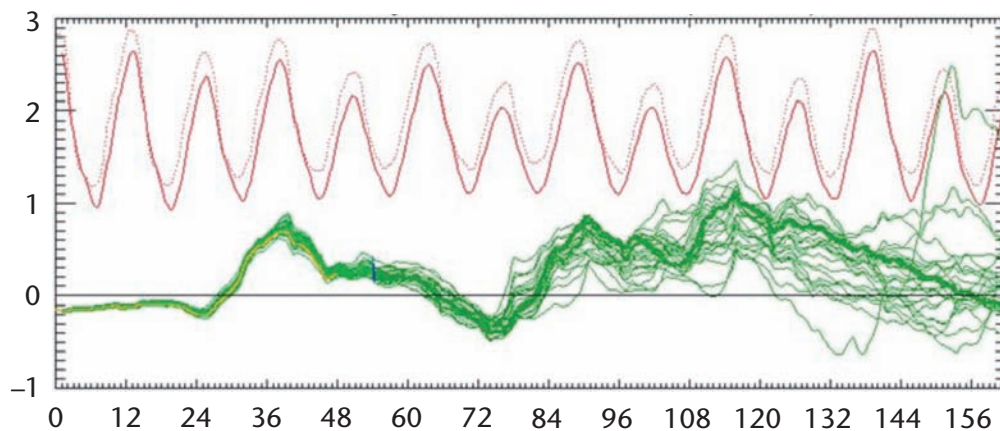
Source: UK Met Office, © British Crown Copyright

Figure 10. Ensemble surge elevation for Lowestoft from 1800 UTC on 2 October 2011 (solid/dashed red lines: alert level at which the surge presents a flood risk, varying according to the tidal level; green/orange lines: ensemble/deterministic surge forecast. Flood risk occurs where the surge exceeds the alert level.)

(b)   If we say there is a 30 per cent probability that more than 10 mm of rain will fall, and the observation shows that only 1 mm fell, the forecast is not right or wrong. We have to measure the actual observed amount for many occasions when we make such a forecast – out of every 100 times that we say this, we should get over 10 mm on 30 occasions. This is what the forecast means. Out of 100 times that we predict 80 per cent probability, we should get it 80 times.

(c)   The simplest way to present verification is using a reliability diagram, which plots a graph of the observed frequency against the forecast frequencies – so it plots exactly the test described above. Three examples of reliability diagrams for probabilities of wind speeds exceeding Beaufort Force 8, 9 and 10 are shown in Figure 11. The ideal is that the line should lie up the main diagonal, from (0,0) to (1,1). The diagram on the left (Force 8) is quite good and shows that forecasts of high probability do mean the event is much more likely – the slope of the graph is slightly less than ideal, but good. The centre diagram (Force 9) is similar but not quite so good for the highest probabilities at the top right of the graph. The diagram on the right (Force 10) shows useful skill for probabilities up to 30 per cent, but at probabilities above that there is no useful information. In fact, this is a rare event and there are not enough samples in the dataset to measure whether there is useful skill. This is a common problem with verifying extreme events; there are not enough data to measure probabilistic skill.

(d)   There are many other measures of probabilistic forecast. Some other common ones are listed here. Much more information is easily available from an Internet search for these terms, or from standard guides to forecast verification.

   (i)   Brier Score: a root mean square error for probability forecasts of a particular event threshold;

   (ii)   Brier Skill Score: compares the Brier Score of the forecasts with the Brier Score of some reference forecast system;

   (iii)   Reliability: measures how well forecast probabilities match observed frequencies;

   (iv)   Resolution: measures how good the system is at predicting probabilities that are different from "normal";

   (v)   Relative Operating Characteristic (ROC): measure how good the forecasts are for decision-making – similar to resolution;

   (vi)   Continuous Ranked Probability Score (CRPS) and Ranked Probability Score (RPS): like a Brier Score for multiple thresholds of the weather variable.
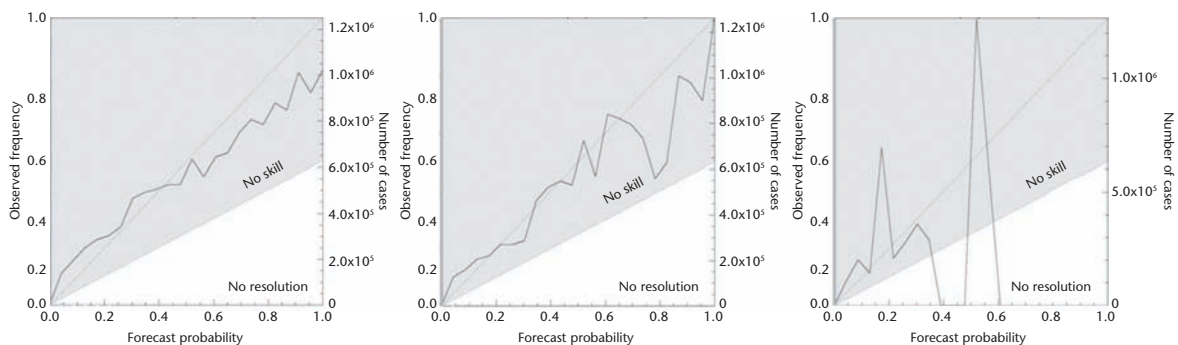
Figure 11. Reliability and sharpness diagram for probabilities of wind speeds exceeding Beaufort Force 8, 9 and 10; T + 36 forecast verification against surface observation over North Atlantic and Europe Model area from January 2006 to February 2007

The WMO Commission for Basic Systems has defined a standard set of verification scores for comparison of EPS, and these are displayed for a number of global EPS at the Lead Centre Website at http://epsv.kishou.go.jp/EPSv/.

13.   FORECASTER TRAINING

In general, forecaster training should include components on predictability and ensemble forecasting, as follows:

(a)   Motivation for probabilistic forecasts – chaos theory and its impact;

(b)   Statistical background theory and approaches;

(c)   Aims of initial condition and model perturbations;

(d)   Standard ensemble verification tools and their meaning;

(e)   Explanation of basic meaning of products (for example, lines on chart);

(f)   Methods of post-processing and their impacts.

Learning Through Doing

The training of forecasters in the use of EPS guidance should be a practical experience using tools that are as close as possible to those used in operations. The optimal benefit from practical training on EPS is only obtained when an NMHS has access to operational EPS data, the operational time to use them and the products and tools to make direct use of them.

The benefits of training that is not reinforced by operational practice are rapidly lost.

(a)   Provision of training in conjunction with a demonstration project such as the SWFDP can help to ensure that the training is reinforced and consolidated by the provision of relevant operational EPS data.

(b)   During training, case studies should be worked through showing the appropriate use of EPS guidance, both in routine and severe weather scenarios.

(c)   Web-based tools can be valuable in training, as they can be used on any workstation system through a standard browser to ensure continued access afterwards.

(d)   In the relatively new area of EPS, periodic training is expected to generate the best benefit. Forecasters require time to build experience in using this guidance followed by further training to reinforce key concepts. It would also be of benefit if various NMHSs could share their experience with the use of EPS.

(e)   Training resources:

    –   *User Guide to ECMWF Forecast Products*
          http://www.ecmwf.int/products/forecasts/guide/

    –   COMET ensemble modules
          https://www.meted.ucar.edu/training_detail.php?orderBy=&topic=15

_____