Use and discussions of the Ensemble forecasts at the Swedish Meteorological and Hydrological Institute

Anders Persson, SMHI

1 Introduction: The January 2005 "Gudrun" hurricane

Saturday 8 January 2005 Denmark, southernmost Sweden and the Baltic region were struck by a damaging storm, "Gudrun"[1], with hurricane winds up to 40 m/s. In total about 20 people died, 7 of whom in Sweden, where the forecast devastation was extensive. In the ensuing clean up operation another dozen people died. To those fatalities must also be counted numerous suicides which took place among those who had lost most or all of their life fortunes due to the forest devastation. "Gudrun" was the worst storm to hit Sweden for 35 years. It brought gloom to a country already stricken by the tsunami catastrophe, where proportionally many Swedes perished. The ECMWF T511 operational forecasts T511, as well as the T255 EPS Control, provided almost perfect short-range forecasts, which enabled SMHI to issue hurricane warnings on Friday 7 January. SMHI would in due course receive an acknowledgement from the Swedish government for these timely and accurate warnings.

However, before Thursday 6 January, the medium range guidance only gave indications about "normally" windy weather. The first attempt by the deterministic T511 and T255 to catch "Gudrun" was in the Thursday 6 January 00 UTC + 60 h forecast where a vortex, with strong gradients although not of hurricane strength, was shown passing eastward over southern Scandinavia. At the 10 April "Gudrun seminar" at SMHI it would have been appropriate to show the audience that the EPS had provided even earlier medium range warnings. But during the days leading up to the storm there were few, if any individual EPS members with hurricane winds.

In the table below are listed the number of EPS members, which forecast very strong MSLP gradients over Denmark or southernmost Sweden, which could be *interpreted* as yielding hurricane winds, together with members with vortices of any significance, including strong vortices forecast too far NW:

	Hurricane, good timing	Hurricane, poor timing	Storm, good timing	Norwegian Sea vortex
3 Jan 12 UTC +120 h		2	1	1
4 Jan 00 UTC +108 h				
4 Jan 12 UTC + 96 h		4		1
5 Jan 00 UTC + 84 h	1	3	2	3
5 Jan 12 UTC + 72 h	2	3	3	5
6 Jan 00 UTC + 60 h	3	2	7	6
6 Jan 12 UTC + 48 h	13	4	6	6
7 Jan 00 UTC + 36 h	11	7	11	11

In the EPS information available on Thursday (6 January 00 UTC + 60 h and earlier runs) only about 4-5 EPS members on average indicated a possible hurricane, most of them with timing errors. Only when the hurricane on Thursday 6 January 12 UTC + 48 hours (operationally available on Friday morning) was perfectly forecast by the T511 and T255 did many EPS members follow suit. Since the deterministic T255 model forecast the hurricane as early and accurately as the T511, the coarser resolution difference does not seem to have played any negative role for the EPS performance.

Were the relatively cautious signals coming from the EPS a sign that the "Gudrun" hurricane was extremely unpredictable, or was it an indication that the EPS had shortcomings? Any forecast system, in particular a probabilistic one, should not be judged just on one case. There were, however, other indications that the disappointing performance of the EPS could not only be explained by "bad luck".

[1] It was at the time named "Gudrun" by the Norwegian Meteorological Service.

2 Scepticism of the EPS

The EPS system was regarded with some scepticism by many meteorologists in Sweden. The forecasters felt that it confused them more than it helped. If this was because the system was deficient or if they were uncomfortable with the novelty of the probabilistic approach, was not quite clear. Equally strong, but different in scope, was the scepticism aired by scientists at SMHI and the Meteorological Institute at Stockholm University (MISU). Their criticism focussed on the perturbation method, which they regarded as unsatisfactory.

Some months after the "Gudrun Seminar" a test to run the HIRLAM six days ahead with boundary conditions from 50 EPS members was made at SMHI. The idea was to see if the combination of EPS and HIRLAM would provide more useful details for a case with severe flooding. It did not, but it came as a surprise that many of the members of this HIRLAM ensemble displayed quite unlikely forecast alternatives already 2-3 days into the forecast[2]. This diversity of solutions did not reflect the forecast inconsistencies, neither with the HIRLAM nor with the T511/T255. The result of the experiment seemed to support the opinion, expressed by the forecasters, that the EPS more confused than enlightened them.

At about the same time, in June 2005, during the ECMWF User's Meeting, it emerged in informal discussions with research scientists that the EPS perturbations were subject to some debate also at ECMWF: – *How much worse (if at all) should a perturbed EPS member be allowed to perform compared to the non-perturbed EPS Control?* The debate touched upon fundamental questions. What is the purpose of the EPS? What constitutes a "good" EPS? Since I had worked at the ECMWF, intermittently as a consultant from 1983, in 1991–2001 as a Staff Member, I have had the fortune to follow and take part in the start and development of the EPS.

3 The development of the EPS 1983-2005

When the ECMWF started to disseminate its deterministic forecasts in 1979 they were superior to other models - but only on average. Periods of good forecast were occasionally interrupted by bad ones. Member State forecasters expressed a need to be told in advance how much the last received ECMWF forecast could be trusted[3].

3.1 The problem of "forecast forecast skill"

The first approach, to take the forecast inconsistency or "jumpiness" as an indication of a priori skill, suggested by E. Kalnay and R.N. Hoffmann in 1983, failed, also when advanced statistical regression analyses were used[4]. By 1990 the computers had become powerful enough to allow scientists to apply dynamic-statistical methods, an idea first suggested by E. S. Epstein in 1969 and further developed by C. E. Leith in 1974. If the forecast failures are due to the errors in the initial analyses, by running a set of forecasts from equally likely, but slightly different analyses, simulating analysis errors, it would be possible to estimate how much analysis errors might affect the forecasts.

It is important to remember that the original rationale for the EPS was not what it became later, *to provide probabilistic forecasts*. The original purpose of the EPS was, in line with the requests from the Member States in the 1980's, to provide *a priori skill measures of the operational deterministic forecast* (the average spread should match the average error). Since the EPS had for computational reasons to be run at a low resolution (T63 until 1996) it was, however, difficult to see how the spread from such a coarse model could indicate the skill of a model with much higher resolution (T213 until 1996).

Using the EPS spread to assess the a priori skill of the deterministic *EPS Control* was not without complications either. The problem here was not only that few forecasters used the EPS Control as a deterministic operational tool, but also that an average perturbed member was slightly less skilful than the unperturbed EPS Control. (We will come back to this issue later).

A third alternative, perhaps the most consistent mathematically as well as operationally, was to make use the fact that the spread around the *Ensemble Mean* should on average match its error. Since the relative skill between EPS Control and an average member did not matter in this case it also had operational advantages. The forecasters

[2] It has been known since the start of the EPS that it was slightly over-spreading during the first 48 hours or so, and that it was recommended not to use it for short range purposes. Still, during recent years some meteorological centres have been conducting experiments using the EPS for providing boundary conditions to limited area models.

[3] Beyond 24 or 36 hours it is not normally possible to identify potentially bad NWPs by comparing it with later observations. Non-linear interactions with upstream systems within an area of influence extending by 30 lon deg per forecast day would make this increasingly difficult.

[4] Later it was found that the simplest way to improve the consistency-skill correlation was to make the forecast system worse. In a forecast system with no anomaly correlation between forecasts and verifying analyses, and between consecutive forecasts, the consistency-skill correlation would reach its maximum value of 0.5 (ECMWF User Guide, Appendix 1).

in the mid 1990's, in particular at the UK Meteorological Office, had begun to discover that the Ensemble Mean tended to be more accurate and much less jumpy than the operational deterministic forecast and the EPS Control.

This discovery logically led to the question if such accurate and consistent Ensemble Mean forecasts were in the same need of a priori EPS skill estimations as the less skilful and "jumpy" deterministic? This could indeed be questioned from a "forecast forecast-skill" perspective, but not from a conviction that the ultimate use of the EPS is to provide probabilities, in particular extreme events, which became more emphasized from the mid-1990's.

3.2 The relation between the probabilistic and deterministic forecast systems

With the introduction in December 1996 of 50 members at T159 (130 km compared to 50 km for the operational T399), it became more and more unattainable that the EPS spread should be used only to estimate the a priori skill of another model. If the Ensemble Mean, in spite of the coarser resolution provided more accurate deterministic forecasts that the higher resolution "state-of-art" model, the question might arise about the relation between the two systems. It became feasible that the EPS could provide forecasts *in its own right* both probabilistic and deterministic (the EM or median). This definitely brought up the question (not to go away) what role the high resolution deterministic forecast would play in relation to EPS? Would the higher resolution model become some Formula-1 for meteorological Michael Schumachers, while the EPS would be a meteorological Alpha Romeo for normal people to buy and drive?

3.3 Probabilities and the quest for multimodality

In our enthusiasm for probabilities we were sometimes overzealous. Early on there was an idea about the "ideal Probability Density Function (PDF)" which the 32 EPS members were supposed to mirror. Soon it was realized that there is no such thing as "PDF of the truth" (except perhaps in quantum mechanics) but just *one* value, the verifying Truth itself.

Our idea about the "ideal PDF" might have had its root in another inspiring hypothesis, this time with more scientific substance. In 1979 Charney-DeVore had presented the idea that the atmosphere was not Gaussian but had a limited number of preferred low-frequency states. When the atmosphere was in one of these states it was more predictable, but became less predictable when it changed from one state to the other. The existence of such multiple atmospheric regimes, if proven, was seen to have had far-reaching consequences for our understanding of the climate system, for the detection and interpretation of climate change and open up the possibilities to forecast the forecast skill. Since Charney died soon after their paper was published and DeVore left meteorology, their hypothesis was left as a meteorological "Fermat's Last Theorem" to be proven or disproven by later generations. Whereas it was enough to have a 10-12 member ensemble to make a fair estimate of the Ensemble Mean and its variance for Gaussian atmosphere, it would need 30 or more EPS members to help us reveal any atmospheric multimodality. To our disappointment there were only rare cases of the ensemble splitting up into two or more discreet flow patterns ("bifurcations").

The quest for the multi-modality influenced how the EPS idea was presented to the forecasters. A common illustration in those early days was a diagram with all the perturbed forecasts starting from one small analysis-PDF but then diverging towards a bi- or tri-modal PDF. Whereas the unperturbed EPS Control in these idealized images always hit the "wrong" mode, a majority of the perturbed members were shown to hit the "right" one. This conveyed a somewhat unrealistic image that a majority of the EPS would unavoidably find the truth.

4 Promoting EPS outside ECMWF

Returning to Sweden with the task of promoting the EPS posed new challenges. In some aspects it was necessary to break with the ECMWF culture in the way the EPS concept and products were presented to the users.

4.1 Interest groups outside the meteorological community

While the EPS information designed at ECMWF was mainly addressed **to forecasters or at least meteorologists**, it turned out that in Sweden also non-meteorologists were interested. Indeed non-meteorologists were even more interested in the ensemble idea and understood it without much problem. Sweden is a country with thousands of lakes and hundreds of rivers and consequently hydrology constitutes an important and substantial part of SMHI, as also the "H" in its name indicates. One of the hydrologists' most important duties is to make run-off forecasts and the EPS possibility to offer rainfall probabilities fitted well into this activity. During the 1990's SMHI had undergone rather profound changes which involved large decentralisation and the development of a strong commercial division led by staff with experience from the private sector. They saw in the EPS a new promising product, suitable for non-manual operational applications. The last couple of years have therefore seen SMHI develop hydrological and meteorological EPS applications, also for commercial use. For details see the SMHI contribution by Mikael Hellgren and Anders Persson as *.pdf or *.ppt at
http://www.ecmwf.int/newsevents/meetings/forecast_products_user/Presentations2005/

4.2 The emphasis on probabilities and categorical forecasts

The EPS information from the ECMWF had ambitiously emphasized **the advantages of expressing forecasts in terms of probabilities**. Unfortunately this did not have any significant impact, and one might wonder why. As history shows, the main force behind changes in weather forecasting comes from the world outside the meteorological community. The invention of the telegraph in the 1830 made distribution of real time observations and forecasts possible, the demand from the growing aircraft industry in the 1920's paved the way for the Bergen school air mass concepts, the enormous expansion of the computer and satellite technology has spurred the development of NWP. Still all these changes took more than 20 years to be accepted. We must realize that establishing probability thinking among decision makers (a challenge not only for meteorologists) will be a slow process. For today's decision makers deterministic forecasts, although inferior to probabilistic, offer the political advantage of relieving any responsibility from their shoulders by allowing them to put any blame on the forecasts.

The main problem the medium range forecasters face today is therefore not a demand of probabilities. Their main problem is not even the occurrence of bad forecasts – but frequent cases of *jumpy forecasts*. It takes five days to realize that a D+5 day forecast is wrong, it takes 12-24 hours to realize that the next medium range forecast has changed abruptly, has made a "U-turn". The EPS Mean, acting as a dynamical filter, offers qualified guidance by *eliminating* smaller and less predictable scales. These removed scales are then *brought back* in a consistent way in the form of probabilities of certain weather events. The use of the Ensemble Mean is therefore in no conflict with the introduction of probabilities, *on the contrary*. The smoothness of the ensemble mean fields rather *invites* the use of supplementary probabilities. It would be detrimental both for the quality of weather forecasts and the development of the EPS to neglect or diminish the usefulness of the Ensemble Mean.

4.3 The importance of statistical verification score to promote the use of EPS

At ECMWF **objective verification statistics** played an important, often decisive role in promoting the use of EPS. The hope was that when the high quality of the EPS was shown to the forecasters, they would start to use it. Experiences from other walks of life tell us that such objective or scientific advice is not enough to change deep rooted habits. We tend to trust statistics only when it confirms our preconceptions. On the other hand, as with physics (including meteorology), where the main difficulty is not the mathematics but how it relates to observations, the problem with statistics lies in the interpretation of the results. Considering the problems of interpreting the *deterministic* verification statistics (see below) it is not surprising that the *probabilistic* scores cause even more confusion.

5 Problems of verification interpretation

The statisticians themselves are the first to tell us that the statistics itself cannot make decisions for us. There are essentially no "objective" verifications – the conclusions are unavoidably subjective: – *Have we succeeded? Do we have a problem? What shall we do?*

5.1 The multitude of verification scores

While the deterministic forecasts by tradition have been verified by a handful of scores (RMSE, ACC, Mean Error, Mean Absolute Error, True Skill Score and Threat Scores), for the EPS we are presented with an additional handful of verification scores (Brier Score, Brier Skill Score, ROC-area, Rank Probability Score and Relative Improvement Index). These are usually applied on 500 hPa geopotential, 850 hPa temperature and rainfall forecasts for Europe, Northern Hemisphere and other areas. We already know that the RMSE and ACC can convey quite different impressions (see the ECMWF User Guide); this is even more true for all the different verification methods applied on the EPS.

Since different scoring system measure different aspects of the forecasts this diversity should not be seen as a problem (or an opportunity to select the most favourable verification) but a useful statistical material to draw conclusions from.

5.2 The mathematical properties of some verification scores

As an illustrative example, take the Mean Square Error (MSE) and decompose it around **c** , the climate value of the verifying day:

$$MSE = \overline{(f-a)^2} = \overline{(f-c)^2} + \overline{(a-c)^2} - 2\overline{(f-c)(a-c)}$$

where **f** is the forecast (assumed without a bias), **a** the analysis (assumed perfect). What we normally associate with forecast skill is described by the third term, the agreement between forecast and observed anomalies. The higher this term the lower the MSE. Normalized by ∣ f-c ∣ and ∣ a-c ∣ it yields the Anomaly Correlation Coefficient (ACC).

But the MSE can also decrease if $\overline{(f-c)^2} < \overline{(a-c)^2}$ i.e. NWP model is unable to realistically simulate the atmospheric variability (explosive cyclones, omega blocking and cut-off lows). A decrease of MSE due to a model deficiency (poor dynamic activity) might make any subsequent genuine improvement of the forecast system difficult to verify, covered as it is by an artificial error reduction. Experience also shows that such "favourable" systematic errors can be coupled to other "unfavourable" systematic errors such as strong biases.

On the other hand, suppression or smoothing of certain unpredictable scales, although "bad" in a NWP system, might be "good" for customer orientated, end-user deterministic forecasts. An experienced forecaster who is aware of what scales cannot be predicted at a certain forecasts range omit these from his forecast, in the same way as the Ensemble Mean acts as a dynamic filter, as mentioned above.

The statistical scores might be “objective”, but whether they are “good” or “bad” remains essentially a subjective matter. Our success and failure cannot be decided by some “index” increasing or decreasing.

5.3 Increased statistical diffusion can deceptively “improve” the Brier score

If we decompose the Brier Score (BS) in the same way as the MSE

$$BS = \overline{(p-o)^2} = \overline{(p-\bar{o})^2} + \overline{(o-\bar{o})^2} - 2\overline{(p-\bar{o})(a-\bar{o})}$$

where **p** is the forecast probability, **o** the event (0 and 1) and $\bar{\mathbf{o}}$ the climatological average[5]. The last term, the agreement between the forecast probabilities and the observed event is often referred to as “reliability”, is not the only factor that can decrease the Brier Score. As with the RMSE, the first term in the BS decomposition, which can be referred to as the “resolution” or “sharpness”, can also contribute to a decrease of the BS if the forecast probabilities are drawn toward the climatological probability, instead of to the outer limits 0% and 100%

It is even more necessary with the EPS than with the deterministic forecast to invest time and effort into interpretation what the statistics actually *mean*, because of the more complex nature of probability forecasting. This will make the statistical verification statistics pedagogically more convincing[6].

6 Possible questions related to today’s EPS

In spite of the progress of operational EPS applications at SMHI in 2004-2005 our scientists were still sceptical of the EPS due to the quality of the perturbations. The disappointing EPS performance of “Gudrun” and the overspread HIRLAM experiment pointed to possible shortcomings.

6.1 Do the perturbed members have to be much worse than the EPS Control?

A high quality ECMWF analysis is variationally optimised and any change is more likely to make it 41% (√2 -1) “worse”[7] and consequently also the perturbed forecasts[8]. In 1992-2000 this deterioration was negligible: 6-12 h lower predictability or 5% lower ACC around D+6. This “√2 effect” was discussed already in 1994 among us who worked with the EPS at the ECMWF, but we regarded the small forecast difference of minor importance. It was assumed that this problem would gradually disappear as the analysis and forecast system improved. In an ideal future, all EPS members would be as skilful as EPS Control.

[5] To emphasise the similarity with the decomposition of the RMSE a slightly different decomposition than suggested by Allan Murphy is applied.

[6] Other ways statistics can be misleading are under-sampling, drawing conclusions from non-representative data and applying a selective choice of period.

[7] See my presentation at the ECMWF workshop at www.ecmwf.int for three different ways to explain why the perturbed analysis error gets 41% worse than the un-perturbed.

[8] The spread and the skill of any ensemble system is not only determined by the size of the perturbations but also their “smartness”, their orientation in phase-space. It must also be remembered that the horizontal scale of the EPS perturbations, T42, (introduced in 1995 and not changed until 1 Feb 2006) is quite large. It corresponds to a spatial scale twelve times larger than the T511, six times larger than T255.

This view is reflected in an important paper by Buizza, Richardson and Palmer from 2003 ("Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's approach", QJRMS 2003, pp. 1269-88) where they presented an inspiring vision of a "perfectly specified EPS" consisting of an infinite number of forecasts, "all equally likely" with the EPS-Control as "a single representative member" of such an ensemble (p. 1283) although they were aware that in the D+2 and D+3 range the EPS Control was generally more skilful than the individual ensemble members (p. 1280).

This was certainly the case in the 1999-2000 material their paper was based on. But obviously something happened since then. During informal discussions at the ECMWF Users' Meeting in June 2005, it emerged that for the last 4-5 years the difference in predictive skill between the EPS Control and an average member has increased and, beyond D+3, now amounts to *1½ to 2 days*. The difference can be seen both in 2 meter temperature point forecasts as well as 500 hPa European or Hemispheric geopotential fields (figures 1 and 2).

What EPS members lacks in individual skill they compensates for in the accumulated information of 50 forecasts. Verifications and operational experiences had shown that the average of the EPS together with the probability information provides high quality information which might be more useful than the T511. It is still an open question if the degradation of the skill of individual EPS members is a necessary sacrifice to obtain good probabilities and good skill scores, or a problem.

6.2 How much has the EPS improved?

Statistics show a good spread skill relation for the EM[9]. But this refers mostly to a good match between the overall (average) spread versus the overall (average) error for individual forecast ranges. This is not the same as a *simultaneous* good match, where situations with small EPS spread yield accurate forecasts and large spread does not exclude accurate forecasts, but mostly lead to less accurate.

In an interesting article in the ECMWF Newsletter in summer 2005, Roberto Buizza showed that the predictability of T255 EPS Control has improved steadily by 1 day since 1994, the EPS by 1½ days. Buizza listed three main causes for the 1994-2005 improvements:

1. **Increases of the model resolutions in 1996 (to T159) and in 2000 (to T255).**
2. **Increase in ensemble size in 1996 from 32+1 to 50+1**
3. **Introduction of evolved singular vectors and of stochastic physics in 1998.**

The first two are not really EPS-related (general model improvements and acquisition of more powerful computers). The third indeed relates to the EPS perturbations but might, due to the nature of the changes, also have introduced some diffusive properties in the EPS.

An examination of all available verifications statistics (provided for example by the 2005 SAC and TAC documentation[10]) shows that the main EPS improvement occurred between 1992-2000, and that the scores after 2001 might have levelled out, some even suggesting no significant improvement during the last five years[11]. The specific EPS contributions have been negative since 2001, but this *negative* trend has been compensated by model improvements to yield rather "flat" scores[12]. For one reason or the other[13], the EPS does not seem to have been able to keep up with the last years' model- and analysis improvements in the T511/T255 models.[14]

[9] With the above mentioned 1½ to 2 days forecast difference such a comparison would be meaningless.

[10] Figure 9 in document TAC35(05)2 p.9 and figure 7 in document SAC(34)2 p.9 do not, as indicated in the main text, refer to EPS but the T511.

[11] Skill scores and ACC with a reference climate that is different to the sample climate might indicate some additional but spurious skill.

[12] Verification statistics of the ACC NH 500 hPa scores of the average perturbed D+5 EPS member has improved by 5%, the average D+7 member only by 2-3%, whereas the EPS Control had improved by 8%.

[13] One explanation relates to a technical bug effective from June 2000 to January 2001 which caused a serious reduction in spread. When the bug was corrected by increasing the size of the perturbations, the spread was obviously set higher than before, with an average ensemble spread alleged to be similar to Control forecast errors at day 5 rather than at day 2 as it was before. The overall probabilities from the ensemble seemed to verify better than before, but it was already then felt that on occasions the spread might be too large.

[14] According to Buizza's article there were some major scientific changes to the EPS in 1998 (such as the introduction of evolved singular vectors and stochastic physics) followed by less profound in 2002 (tropical SV) and 2004 (sampling strategy). The increased vertical and horizontal resolutions in 1999-2000 were no specific EPS implementations.

6.3 *Recent synoptic examples of the EPS versus the deterministic forecasts*

It is important that the EPS is validated not only from statistical verification scores, but also from numerous operational case studies. Since December 2005 the synoptic behaviour of the EPS has been monitored, in particular in connection with cases with extreme weather events. Such events occurred on 12 December with hurricane winds over N Scandinavia, on 16 December when a small-scale storm hit Northern Germany and 11-12 January when a cyclone with hurricane winds moved from Scotland to Northern Scandinavia. The impression of this limited number of cases is that the EPS is slow to warn about extreme events compared to the T511/T255. Once the deterministic forecasts 5-7 days in advance have consistently indicated a certain weather event, it takes the EPS some days more to confirm the event. In the meantime it keeps warning about other possible and, in their dynamic context, impossible synoptic scenarios. An exception occurred on 17 Dec 00 UTC when a majority of EPS-members, in conflict with both the T511 and their own T255 EPS Control, forecast mild weather, which verified.

7 Advantages and disadvantages with lagged average forecasting

With a difference in predictive skill of 36-48 hours between an average perturbed member and the EPS Control (perhaps 48 hours in comparison with T511) there is a possibility over 48 hour time period to receive five T511 deterministic forecasts which have a higher or equal quality as an average EPS member. While we continue to run and make use of the current EPS applications, at SMHI we currently explore the use of lagged T511 both for medium range and in any HIRLAM-EPS application. The advantages with lagged T511 forecasts are:

1 High resolution throughout the 10-day forecast, no problems with accommodating a deterministic forecast with a higher resolution

2 The lagged averaged forecast can be used from the start of the forecast, whereas the EPS is suitable only from 2-3 days into the forecast

3 The lagged average provides, thanks to the smoothing, more accurate categorical forecasts with reduced jumpiness

4 The lagged approach is more economic because forecasts data need less storage space

Disadvantages is a slightly higher degree of jumpiness than EPS, including its Ensemble Mean. Equal weights on the lagged members will result in coarser probability intervals (20-25% intervals for 4-5 members instead of 2% for the EPS). However, if the weights are *non-equal* (proportional to the average forecast quality), the probability intervals might, at least formally become much smaller[15].

8 Summary and recommendations

1 *What purpose should the EPS fulfil?* Estimating the skill of a higher resolution deterministic model or constitute an independent forecast system? If the latter is the case, should it estimate the a priori skill of the non-perturbed model or the ensemble mean?

2 *What is the relation between probabilities and any deterministic forecast?* Only the ensemble system can provide probabilities, but they must necessarily have a consistent relation to any deterministic forecasts values. Should these be taken from the operational model, the EPS Control or the Ensemble Mean?

3 *The perturbation technique needs to be re-considered.* There is nothing to be done about the "√2 - effect", but the efficiency of the perturbations is not only determined by their amplitude or geographical extension. The use of perturbation for monthly forecasts and seasonal forecasts should also be considered. Would more frequent, deterministic forecasts with higher resolution in a lagged mode provide better forecasts?

4 *What constitutes a good probabilistic system?* Any probabilistic system must have a high reliability; what makes the difference in predictive skill is also a matter of the degree of sharpness, resolution. Good reliability is only a necessary, but not sufficient condition for a good EPS. The skill of the EPS should not only be measured in reference to a single deterministic forecast but with a cleverly built lagged ensemble.

5 *The daily monitoring must be improved.* The EPS must be subject to a more elaborate statistical analysis, where the shortcomings, limitations or other mathematical artefacts of the different scoring systems must be taken into account.

Both scientists and forecasters in Sweden have confidence in the EPS approach as such, which we see as the ultimate method to account for the uncertainty in weather forecasting. We hope that the issues raised above can be better understood and the system, one way or the other, improved.

[15] So would for example the weights 10, 20, 30 and 40% applied on four forecasts of different quality (D+4, D+3, D+2 and D+1) in different combinations yield 10% probability intervals from 0% to 100%. A 60% probability would be produced if either D+2 and D+3 indicated the event, as well as D+1, D+2 and D+3.

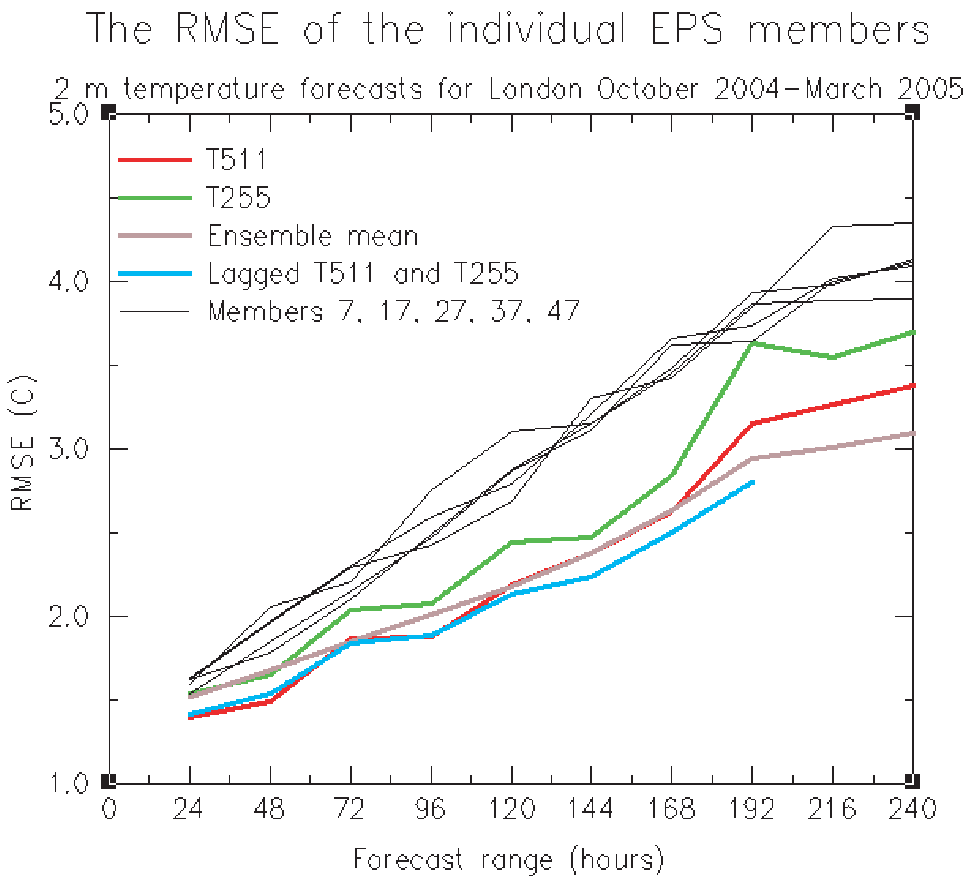

Fig. 1 The RMSE of the 2 meter temperature forecasts for Heathrow October 2004-March 2005. See legend for type of forecast. The difference between the T511 and the T255 is partly due to influences of seapoints in the English Channel in the coarser T255 model and does not truly reflect the forecast quality between the two models. The light blue lagged ensemble mean is a weighted average of the last three days' T511 and T255. Its errors are comparable with the errors of the proper ensemble mean.

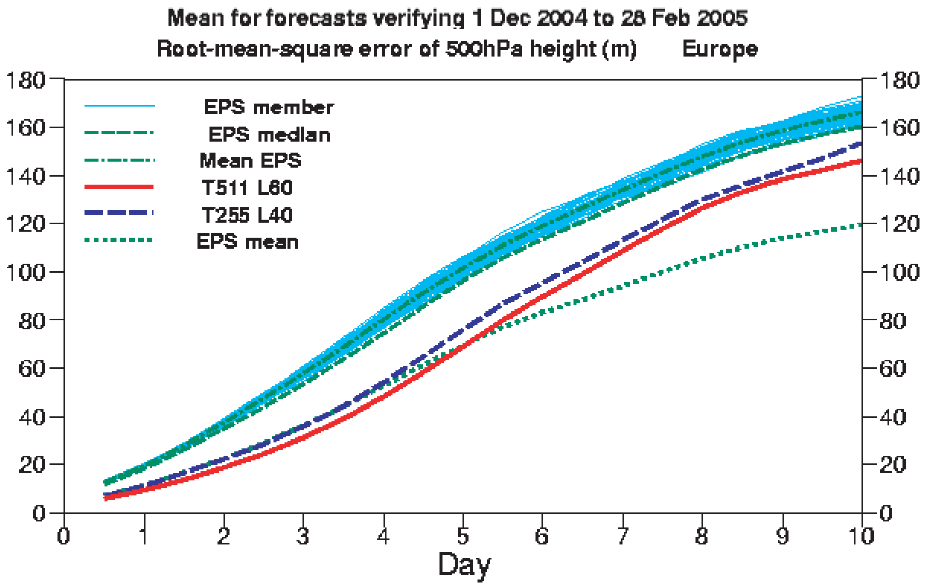

Fig. 2 The RMSE of the 500 hPa ECMWF geoptential forecasts December 2004-February 2006. The T511 has slightly lower RMSE than the unperturbed T255. Note that the EPS Mean has lower errors than T511 after five days. This is also true for the European area (not shown) but for the North American area (not shown) the EPS mean has more or less the same error level as the T511 because the difference in predictive skill between perturbed and non-perturbed members is 2 days.