Problems and prospects in the operational use of medium-range NWP products

by Anders Persson

Introduction: A forecast of forecast skill

20 years ago one of the pioneers in NWP, Bo Ragnar Döös in a lecture "Status and Outlook of Numerical Weather Prediction" summarized the progress over the previous 20 years in words that might also catch today's feelings:

> "After the comparatively good results which were obtained with the simple barotropic model in the early 50:ies one became quite optimistic that the forecast problem could be solved within short. The expectations were really great.
>
> Since that time we certainly have made significant improvements and we can now make more accurate and more detailed forecasts. Nevertheless, we must also admit that the development has not been fast as was expected. It seems that our ability to predict the motions of the atmosphere is levelling off on a plateau..." (Döös, 1971)

The gloom in Döös' "forecast of forecast skill" is understandable. After large initial improvements in the late 50:ies, the skill of the NWP had by 1971 flattened out and during the following years progress was meagre.

In 1979 the advent of ECMWF operational medium range weather forecasts further stimulated the development. In 1984, after five years, ECMWF produced +72 hour 500 hPa forecasts of the same skill as +36 h 500 hPa forecasts at the time of Döös' lecture and +120h forecasts of the same skill as in the 50:ies (fig. 1)

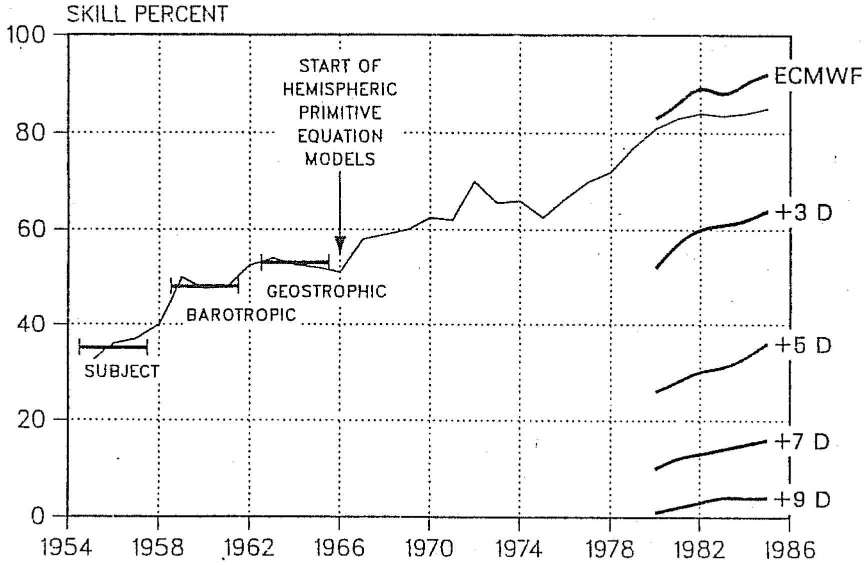


Fig. 1 Record of annual skill of the NMC+36h, 500 hPa forecasts over North America (after Shuman, 1978). Added are the ECMWF scores for D+1, D+3, D+5, D+7 and D+9 for the same area (from ECMWF User Guide, 1988).

In contrast to 1971, the expectations in 1984 were really great and there were hopes that the period of useful deterministic forecasts could be extended towards 12 days within the next ten years. We would by now (1991) have reached almost D+10, but we all know that the true figure is "only" D+6 or D+7. In spite of significant improvements we must once again admit that the development has not been fast as was expected. There are even feelings around that our ability to predict the motions of the atmosphere might be levelling off on a plateau. Hopefully we might this time be too pessimistic. The new T213 model offers a sound numerical foundation, upon which improvements in the analysis and physical parametrization can lead to increased predictive skill.

2. The usefulness of the forecasts

In spite of the proclaimed D+6.5 skill many forecasters are hesitant to use NWP forecasts beyond D+3. Partly this lack of confidence is a result of a tendency to over-interpret the forecasts of synoptic details in the medium range. It is important to be aware of how much skill can be expected for a specific lead time. As a guidance I quote the ECMWF User Guide recommendations from 1986 and the current (1991) American Meteorological Society recommendation:

The Level of useful predictive skill

	1986 ECMWF User Guide	1991 AMS Policy statement
3-5 days:	Forecasted positions of cyclones and frontal systems may be in error Formation of blocks are forecasted with great skill. Near surface parameters have skill as daily means.	Large-scale circulation events (major storms and cold waves) can usually be anticipated. Forecasts of daily temperature have good skill at D+3, fair at D+5; forecasts of rain have fair skill at D+3, marginal at D+5.
5-7 days:	Forecasts of long waves, maintenance and breakdown of blocks fairly reliable. Near surface parameters have skill as mean values over several days.	Daily maximum temperature can be forecasted with modest skill. Mean temperature and total precipitation 6-10 days ahead can be predicted with some skill.
8-10 days:	The model is able to identify periods of higher and lower atmospheric activity, but not the geographical location. No skill in near surface parameters.	

In this context I would also like to recommend the use of spatially smoothed forecast fields (Persson, 1984).

3. Inconsistency

Not all problems in interpreting forecasts in the medium range can be tackled by smoothing the small scale or even the synoptic features. Quite frequently severe uncertainties are experienced also in the large-scale flow pattern: one day a huge blocking is forecasted at D+8, next day a strong zonal at D+7, next day at D+6 a deep stationary trough. During the last years there has been increasing concerns among the Member States about such disturbing inconsistency or "jumpiness" in the ECMWF forecasts.

When the forecaster uses different NWP models he favours the one which generally performs best. Twenty years ago the forecasters also used to compare the NWP with the latest available subjective analysis. The model which had best caught the +24 hour development was expected to be the most skilful at +72 hours (the ultimate limit of predictability in those days).

This approach worked well since the one day error has some linear bearing on the skill of the NWP forecast up to D+3. Nowadays this approach does not work; it is very difficult to spot significant errors in a +24 hour forecast and any errors at D+1 do not show any linear relation to the errors at D+5 and beyond.

During the 80:ies the practice gradually changed. With no other alternative the forecasters were left to look at the consistency between different models and/or between different forecasts from the same model as the only way to express any opinion about the possible skill of the current forecast.

Investigations of the inconsistency, measured as the differences between forecasts verifying at the same time, have shown that the T106 model did not seem to be more inconsistent than other NWP models, nor has it increased during the last years. It also showed that models with lower inconsistency were not necessarily more skilful.

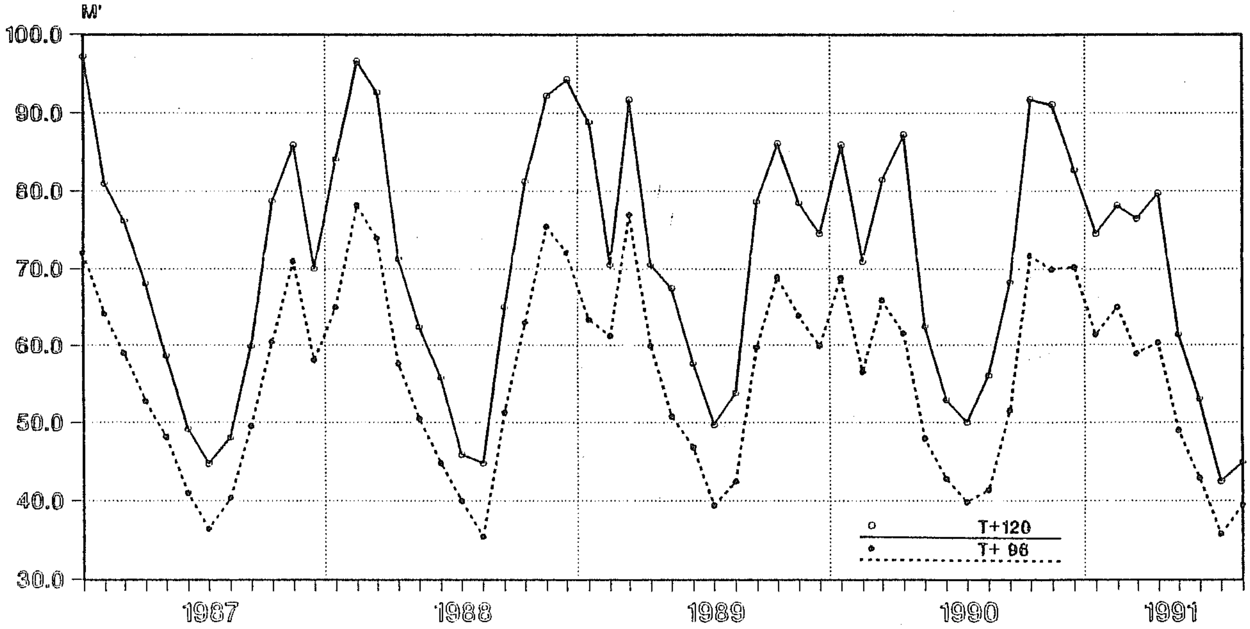


Fig. 2 Difference between D+5/D+4 for ECMWF 500 hPa forecasts for Europe 1987-91.

But is high consistency necessarily something "good" and inconsistency necessarily something "bad"? As much as we want a good D+7 forecast to be "followed up" in next day's D+6, we do not want a D+7 to be followed up if it is bad. Consistency is subordinated to skill. Emphasizing consistency can lead to completely wrong assessments. One example:

Imagine a situation when an improvment of the model has great impact on the D+6 forecasts, less on the D+5 or D+7. This will of course increase the occasions when, on one hand, yesterday's D+6, due to its increased quality, is quite similar to today's D+5; on the other hand also increase the occasions when today's D+6 is quite different from yesterday's generally worse D+7. Forecasters using consistency as a measure of a priori skill will erroneously get the impression that D+5 has improved (because it appears more consistent with yesterday's D+6) and D+6 has become worse (because it is less consistent with yesterday's D+7).

We must realize that we need some degree of inconsistency to allow the model to deviate from yesterday's normally less skilful solutions. But contrast to skill, which ultimate aim is to reduce the error to zero (or the correlations to 100%) there exists no desired value for consistency. Thus consistency can not be used as a tuning parameter, as e.g. the model's climatological behaviour. To understand the effect of inconsistency and skill the typical variability of the atmosphere becomes important.

4. Variability

Consider a forecasting system issuing forecasts on a regular basis. Let us denote them by **F**, **G**, **H**.... and let indices 0, 1, 2, 3, 4... denote the lead time with 0 as the initial analysis. Skill of the forecasts are measured as the relation between F_1 and G_0, F_2 and H_0 etc., the consistency(**c**) as the relation F_2 and G_1, F_3 and G_2, F_4 and H_2, G_2 and H_1 etc. and variability(**v**) by the change from F_0 to F_1, F_1 to F_2, G_2 to G_3 etc. which should agree as closely as possible with the observed general change(**v**). Three consecutive forecasts can be illustrated as in fig. 3.

$$\begin{tikzcd}
\mathbf{F}_0 \arrow[r, "\mathbf{v}_1"] \arrow[dr, "\mathbf{v}"'] & \mathbf{F}_1 \arrow[r, "\mathbf{v}_2"] \arrow[d, dash, "\mathbf{c}_1"] & \mathbf{F}_2 \arrow[r, "\mathbf{v}_3"] \arrow[d, dash, "\mathbf{c}_2"] & \mathbf{F}_3 \arrow[r, "\mathbf{v}_4"] \arrow[d, dash, "\mathbf{c}_3"] & \mathbf{F}_4 \arrow[r] \arrow[d, dash, "\mathbf{c}_4"] & {} \\
 & \mathbf{G}_0 \arrow[r, "\mathbf{v}_1"] \arrow[dr, "\mathbf{v}"'] & \mathbf{G}_1 \arrow[r, "\mathbf{v}_2"] \arrow[d, dash, "\mathbf{c}_1"] & \mathbf{G}_2 \arrow[r, "\mathbf{v}_3"] \arrow[d, dash, "\mathbf{c}_2"] & \mathbf{G}_3 \arrow[r] \arrow[d, dash, "\mathbf{c}_3"] & {} \\
 & & \mathbf{H}_0 \arrow[r, "\mathbf{v}_1"] \arrow[dr] & \mathbf{H}_1 \arrow[r, "\mathbf{v}_2"] & \mathbf{H}_2 \arrow[r] & {} \\
 & & & {} & &
\end{tikzcd}$$

Fig. 3 Schematic illustration of skill, consistence and variability in three consecutive forecasts.

If the variability **v** is slightly larger in the model than in the atmosphere, this will increase the inconsistency **c** and decrease the skill. On the other hand: if the variability is slightly less than in the atmosphere, it will decrease the inconsistency and slightly improve the performance. This is because any predictive system gains from a certain overestimation of the persistence when the decline in skill is larger than the persistance.

Already in 1982 it was noticed at ECMWF (Böttger and Grönaas, 1982) that a D+4½ or D+5 forecast, verified as if it was a D+6, actually scored better than the real D+6.[1] Saha and van den Dool(1988) have developed these ideas further and suggested that the lead time **t** when an **D+t** forecast verified as a **D+t+1** forecasts scores better than the real **D+t+1**, should be defined as objective measure of the practical limit of predictability.

Measurements of the persistence in the ECMWF model shows that the old T106 model was slightly more persistent than the atmosphere, whereas the new T213 model is slightly less. This means that some of the inconsistency reported is unnecessary, caused by the model's over activity. By reducing this over activity the inconsistency can be reduced and the model's skill increased.
But an other consequence of Saha/van den Doll's reasoning is that the increased realism in the new model might not show up in the scores, because T213 is able to develop intense systems during **all** lead times and thus take greater risks in the medium range, whereas the old T106 was more cautious - but also more unrealistic - after D+5.

But to keep an atmospheric model more "sluggish" just to make it perform slightly better in RMSE terms would be in contradiction with our aim to make a model as realistic as possible. One way, and perhaps the only way, to overcome this contradiction, and also to find new tools in medium range weather forecasting, is to develop the **ensemble** or **Monte Carlo** forecast technique.

5. Ensemble forecasting

In the so called ensemble or Monte Carlo forecast technique a large number (20-200) of identical forecast models are run with slightly different initial analyses, reflecting the uncertainty of the initial state. The spread of the resulting forecasts will not only reflect how sensitive the atmosphere at each particular day is to uncertainties in the initial analysis and form a basis for a general "forecast of the forecast skill", but also provide the forecaster with a wide range of forecast alternatives and the probability of different evolutions[2].

[1] This of course reflects the standard of 1982. Today it might be true for a D+6½ or D+7 verified as a D+8 forecast.

[2] The Monte Carlo method was originally introduced by J.v.Neumann and S.M.Ulam around 1945 as a way of solving deterministic mathematical problems using random numbers, either by direct simulation of physical or statistical problems or by reformulating deterministic problems to involve random processes.

As in the mathematical Monte Carlo simulation we will encounter the problem of how to define randomness? How to avoid over exhausting the computer capacity? Even modest calculations have indicated that a complete Monte Carlo system would demand approx. 10^{1500} parallel forecasts!

Apart from the sheer impossibility to run even a fraction of such forecasts (in the future 50, maybe 100 parallel forecasts will be run, never more than 200) come specific meteorological problems: random noise added to the initial state does not realistically interact with the relevant meteorological scales. Already in the initialization or in the early stages of the forecast will random perturbations be smoothed out in the model.

At ECMWF positive results have been obtained in experiments using "fastest growing" perturbations, i.e. error patterns whose amplitude and geographical position make them influence the whole 10-day forecast period. This technique, however, has the disadvantage of violating the principle of randomness in the choice of perturbations, unless a specific probability can be ascribed to each chosen random perturbation. Since the magnitude of these "fastest growing modes" tend to be larger than the normal analysis errors, the technique also increases the risk of over estimating the forecast errors, which also has been seen in some experiments.

6. The operational use of ensemble forecasts

To both the operational forecaster and the end users, the provision of ensemble and probability forecasts will provide information of a type not previously encountered. Individual charts for each forecast alternative can not any longer be seen as the main way to make the forecaster familiar with the forecast alternatives. To cope with the huge amount of data, new forms of presentation will be required to provide users with ready access to essential information available within ensembles e.g.:

a) ensemble mean fields (of temperature, geopotential, upper air winds etc.) where the induced smoothing will provide forecast fields with unpredictable scales filtered out.

b) grouping the ensemble solutions into a few (2-4) distinct regimes using some clustering technique and/or identification of extreme events in any of the ensemble members (e.g. blockings in the mid latitudes or hurricanes in the tropics).

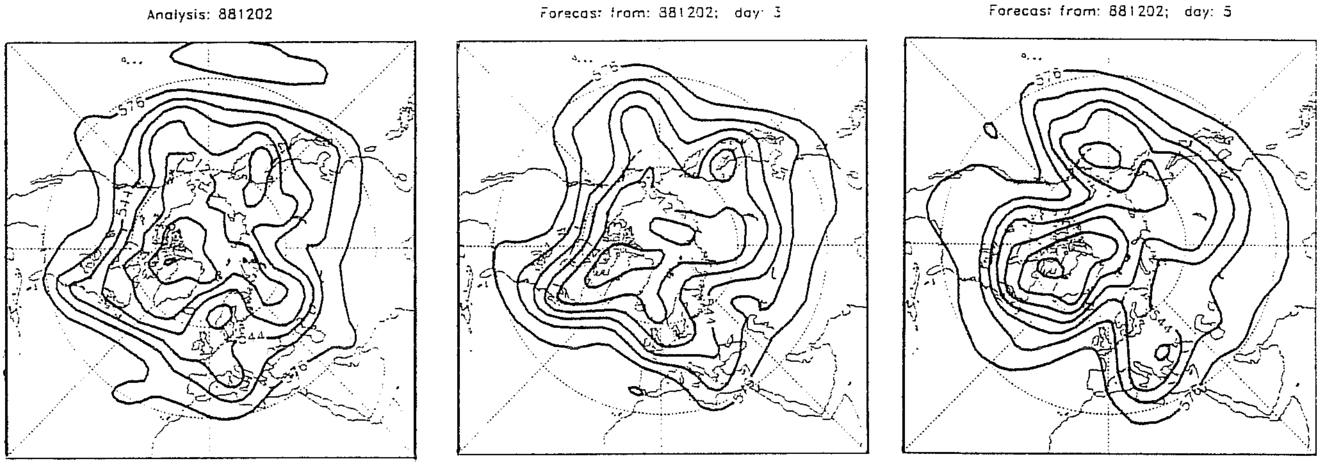
c) frequency distributions of near surface parameters (2M temperature, 10M winds, precipitation, clouds etc) which can serve as a guide for operational useful probability forecasts.

In the course of development we will encounter model output in ways never seen before. This will force us to identify systematic errors such as impossibly high or low variances, non-zero probabilities to weather events never observed etc. As with any new technique thorough assessment of all aspects of ensemble predictions will be required prior to general acceptance of ensembles within the meteorological communities.

To suggest the constraints the huge amount of data will lay on the operational use of the Monte Carlo products, but also how it opens up new possibilities for the forecaster, an experiment from 2 December 1987 will be presented below.

7. A synoptical example

On 2 December 1988 00 UTC the 500 hPa flow pattern was generally blocked over Europe (fig.4a). The operational ECMWF forecast indicated that during the subsequent days the block would break down, a trough move in from the Atlantic (fig. 4b) and develop over SE Europe as a main trough. A new ridge would follow suite and amplify over W Europe (Fig. 4c)



Analysis: 881202

Forecast from: 881202; day: 3

Forecast from: 881202; day: 5

Fig. 4 a) 500 hPa hemispheric 500 hPa analysis 2 December 1988 12 UTC, b) the ECMWF D+3 and the c) ECMWF D+5 forecast.

If a forecaster, by some forecast-forecast skill system, can determine in advance that this forecast is of high quality, then there are no problems; he can use the material in full confidence.

If he cannot determine the predictive skill or if the forecast-forecast skill system tells him the forecast is bad, then he is keen to know the erroneous parts of the forecast and/or which alternative developments are likely: forecasters in Switzerland would be interested in the timing of the trough passing over them around D+3. Forecasters in Greece would like to know the intensity of the low and deepening on D+5; forecasters in Tromsö would need some indications whether the mild Atlantic air might reach northern Norway on D+5 etc. The Monte Carlo technique will provide them with this option.

In one Monte Carlo experiment, 2 December 1988, 24 forecasts with slightly different initial conditions were launched. In fig. 5 a-e only two of these ensemble members (IW1 and IX4) are shown.

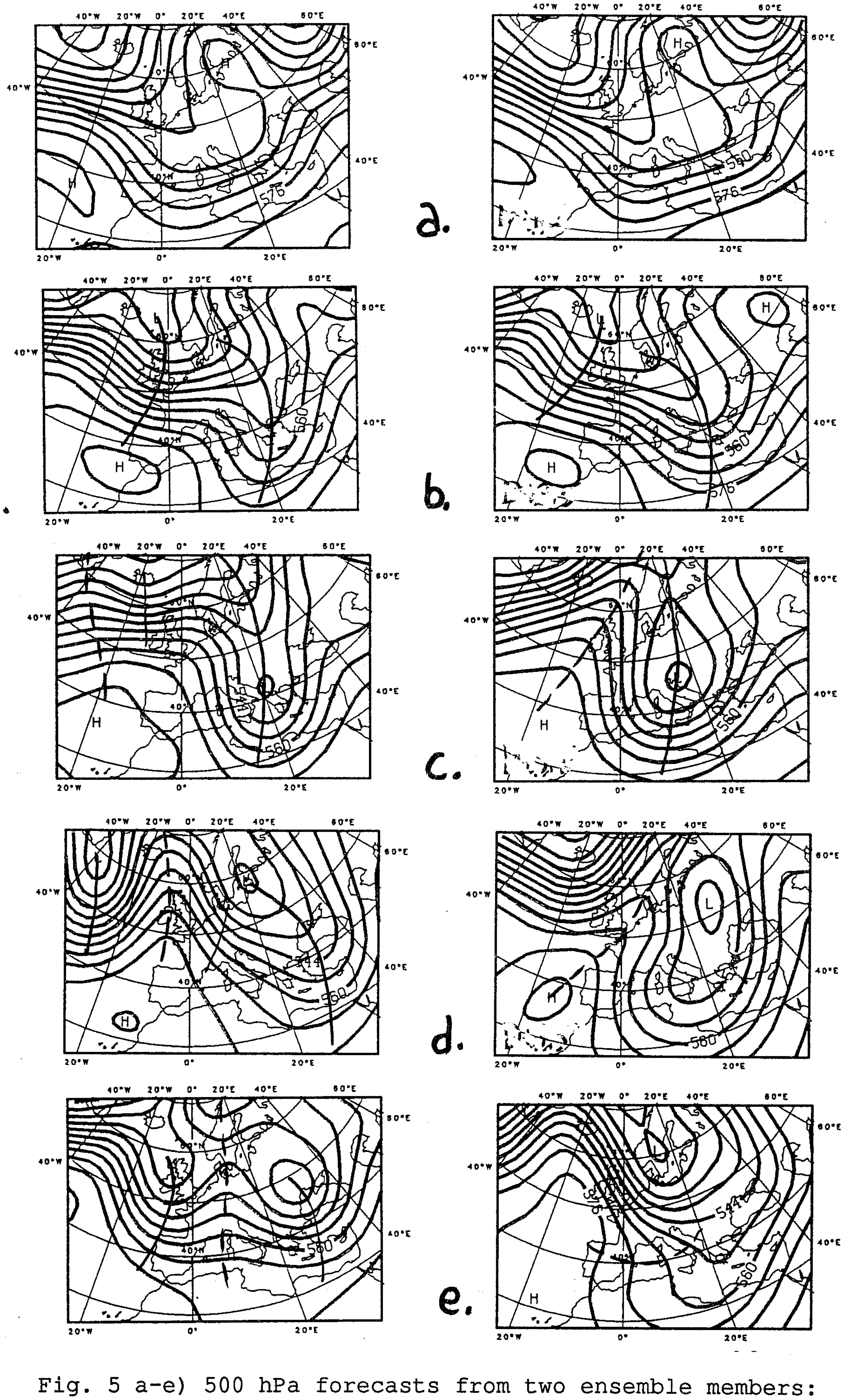


Fig. 5 a-e) 500 hPa forecasts from two ensemble members: b) D+1.5, c) D+3.5, d) D+5.5, e) D+7.5 and f) D+9.5.(From Palmer, Molteni and Mureau, 1990).

8. Probability charts

To be able to assess this huge amount of forecast data, new ways of displaying are necessary, e.g. probability charts of the precipitation (fig.6 a,b)

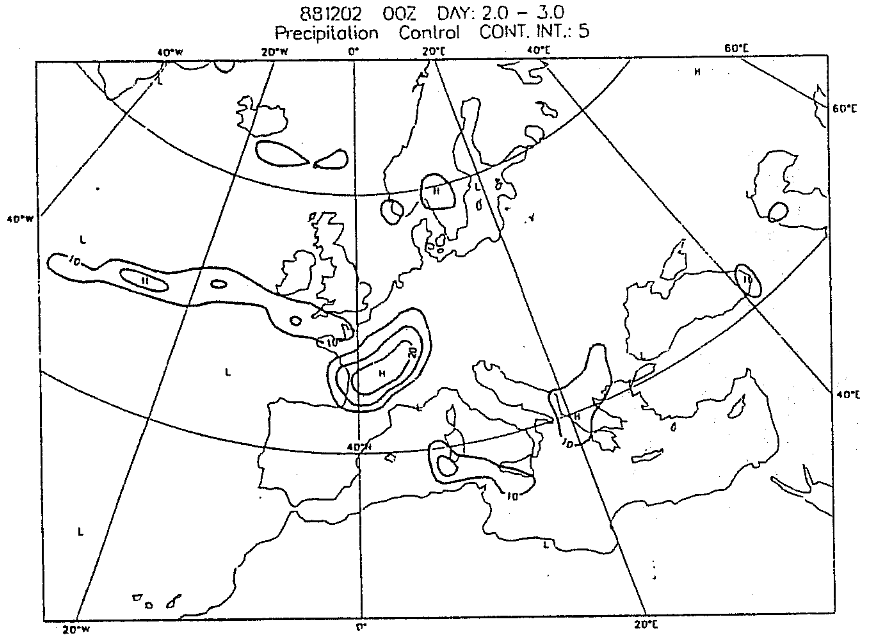

881202 00Z DAY: 2.0 - 3.0
Precipitation Control CONT. INT.: 5
40°W
20°W
0°
20°E
40°E
60°E
60°E
40°E
40°W
20°W
0°
20°E
H
L
10
20

Fig. 6a Deterministic forecast from 2 December 1988 of rainfall accumulated between D+2 and D+3. Contours every 5 mm.

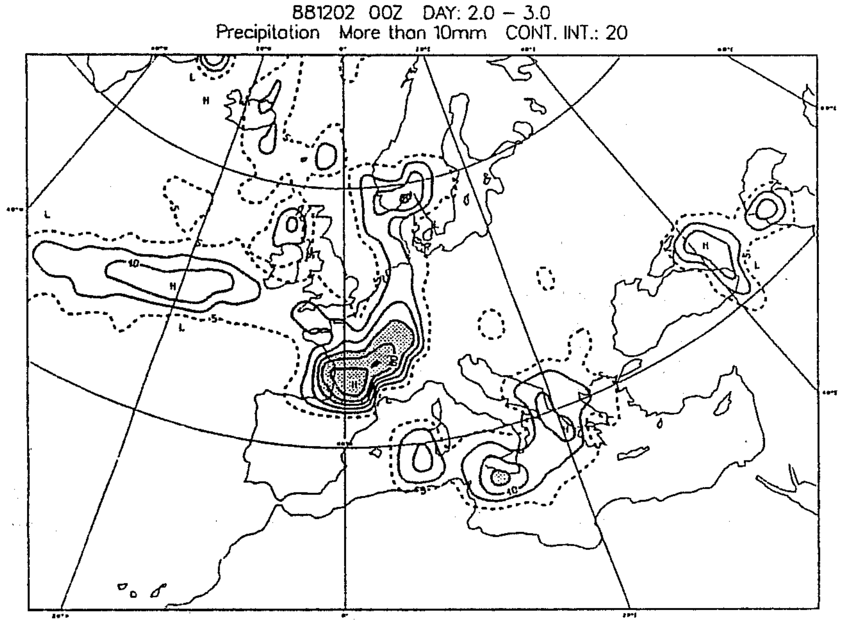

881202 00Z DAY: 2.0 - 3.0
Precipitation More than 10mm CONT. INT.: 20
H
L
5
10

Fig. 6b Probabilistic forecast based on an ensemble of 25 forecasts from 2 December 1988, showing the probability that the amount of rainfall accumulated between D+2 and D+3, will exceed 10 mm. Contours 5, 20, 40, 60, 80 and 100 %. Stripped area indicates probability greater than 60%. (Palmer et.al. 1990)

9. Frequency diagrams

It is also possible to present the forecast of the near surface weather parameters in a frequency diagram, displaying the probability of rainfall amounts or daily temperature (fig.7 a,b):

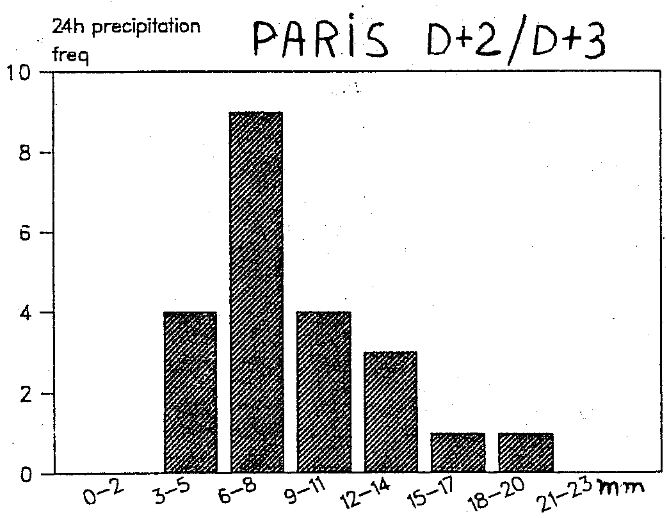


Fig. 7a Frequency diagram of 24 forecasted 24-hour accumulated rainfall amounts in Paris from 2 December 1988 12 UTC +48/+72. Note that the 50-50% value is around 8 mm.

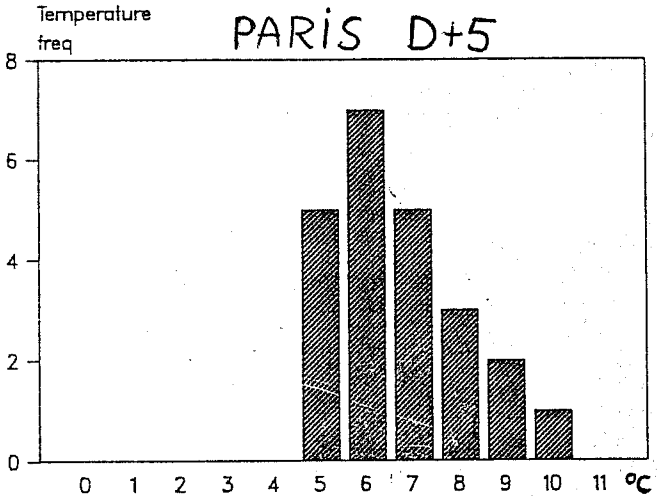


Fig. 7b Frequency diagram of 24 forecasted daily temperature in Paris from 2 December 1988 12 UTC +120, valid 7 December. Note that the 50-50% value is around 6.5 degrees.

(The temperature in Paris reached 7 degrees on 7 December, though milder air was not far away over the Atlantic, a justification for the probabilities of temperatures up to 10 deg.)

In some locations the forecasts of near surface parameters suffer from systematic errors and cannot be used without some statistical correction, e.g. by Kalman filtering (Persson, 1991).

10. Summary

It is too early to say that we for the medium range have reached the ultimate level of deterministic forecast skill. Improvements in observations, analyses and physical parametrization may still contribute to future increases the deterministic usefulness (as reflected in the traditional RMSE and anomaly correlation scores). But there is a growing feeling that the medium range forecast problem, like all forecast problems (not only meteorological), should be dealt with in probabilistic terms.

In a way this has for a long time been intuitively felt by the forecasters. Guided by their experience and the demands from the general public many have come to realize that the practical usefulness of a NWP model is not restricted to the deterministic skill. During the last 10 years several weather services have made use the ECMWF deterministic forecasts to assess probabilities of different weather development, ssubjectively or objectively, qualitatively or quantitatively, though primarily in the small scale.

To assess the probabilities of alternative developments in the large scale, different "poor man's Monte-Carlo methods" have been tried such as looking at different forecasts (from previous days and/or other models). With the advent of an operational Monte Carlo or ensemble forecast system in a couple of years' time, the forecasters will be able to fully explore the usefulness of a truly realistic atmospheric forecast system.

References:

Döös, B.R., 1971: The Staus and Outlook of Numerical Weather Prediction, Zeitschrift für Meteorologie, Band 22, Heft 1-5, 46-53.

Böttger, H. and S. Grønaas, 1982: Optimal verification of ECMWF surface temperature forecasts at two locations in Europe. ECMWF Technical Memorandum No. 64, July 1982.

Palmer, T.N., F. Molteni and R. Mureau, 1990: The Monte Carlo Forecast. Weather, 45, 198-207.

Persson A., 1984: The application of filtered forecast fields ro synoptic weather prediction - presentation of the products and recommendations for their use. ECMWF Technical Memorandum No 95.

Persson, A., 1991: Kalmanfiltering - A new approch to adaptive statistical interpretation of numerical meteorological forecasts, (in: Lectures presented at the WMO Training Workshop on the Interpretation of NWP Products in Terms of Local Weather Phenomena and Their Verification, PSMP Report Series, No 34, WMO/TD No. 421)

Saha, S. and H.M. van den Dool, 1988; A Measure of the Practical Limit of Predictability. Monthl. Wea. Review, 114, 2522-26.

Shuman F.G., 1978: Numerical weather prediction. Bull. Am. Met. Soc. vol. 59; 5-17.