Tellus (1994), 46A, 52–65
Printed in Belgium – all rights reserved

TELLUS
ISSN 0280-6495

Formulation of gaussian probability forecasts based on model extended-range integrations

By M. DÉQUÉ*, J. F. ROYER and R. STROE†, *METEO-FRANCE, Centre National de Recherche Météorologique, 42 Av. Coriolis, Toulouse, France*

(Manuscript received 27 November 1992; in final form 26 July 1993)

ABSTRACT

A sample of 40, 44-day winter forecasts is used to investigate the predictability of 850 hPa temperature over Europe. These forecasts exhibit a significant skill when averages of day 5 to day 14 and day 15 to day 44 are considered. This skill is, however, very close to that of the trivial climatology forecast. A probability forecast is performed, using a gaussian density with the deterministic forecast for the mean, and the climatological standard deviation (SD). The rank probability score (RPS) of such a forecast is better than, but again very close to, that of the probabilistic climatology forecast. The categorical forecast is also studied as a limit case when the SD is zero. The RPS is minimal when using the conditional probabilities of the verification analyses, but the results are not widely improved when robust estimates are used. The results could be widely improved if we used a suitable SD in our forecasts. However, the attempts to predict a priori the optimal SD lead to non-significant results. The best available probability forecast, in our local gaussian approach, uses, for the mean, the regression of the verification analyses by the model forecasts, and, for the SD, a scaled climatological SD.

1. Introduction

The recent numerical weather prediction models have shown that some skill exists beyond the medium-range (Tracton et al., 1989; Sirutis and Miyakoda, 1990; Palmer et al., 1990; Milton and Richardson, 1991; Yamada et al., 1991). However, in the output of the long range forecast, the potentially valuable information is hidden under a great amount of noise, so that it cannot be used in the same way as in short range forecasts. Several methods can be applied to increase the signal to noise ratio. Optimal statistical filtering as in Déqué (1988a) cannot be applied yet, since it would require hundreds of independent forecasts; the presently available datasets are generally not greater than several tens for 30 day forecasts or beyond. The most common way of filtering is time averaging. In this paper we use averages over the days 15–44 of the integration. In this case, we do not try to predict the day-to-day evolution, but only the monthly mean features. Another possible filtering is space averaging, which suppresses the local details. When applied to monthly means, this spatial average does not seem to increase the skill by a large amount (Shea and Madden, 1990). A more expensive filtering consists of ensemble averaging, where each forecast is based on several model integrations. In the Monte Carlo technique (Leith, 1974), the initial conditions are obtained by adding small random perturbations to the initial analysis and are supposed to simulate the observation errors. In the lagged average technique (Hoffman and Kalnay, 1983), the initial conditions are analyses lagged by a small time interval (6 or 12 h): the initial differences are the model short-range errors. The latter technique is frequently used because it is simple to carry out operationally, but it does not allow us to produce large ensembles. However, if we consider monthly mean forecasts, the improvement of skill brought

* Corresponding author.
† Deceased 18 November 1992.

about by ensemble averaging is rapidly saturated with a small number of individual integrations (Déqué, 1991a).

Even though the signal to noise ratio is generally increased with the above techniques, the mean skill remains small, and large fluctuations of skill are observed from one case to another. This has led to investigating the possibility of forecasting the forecast skill. Unfortunately, to our knowledge, no serious predictor has yet been found in the extended range, although some relations between skill and ensemble spread (Kalnay and Dalcher, 1987), or between skill and the Pacific-North America index (Palmer and Tibaldi, 1988) have been found for the medium range.

In fact, trying to predict the forecast skill is a way to represent our confidence in the model forecast. A probabilistic formulation of the forecast is another, more systematic, way to accomplish this: a large confidence in the forecast will correspond to a small variance of the distribution. This latter approach is more general, as far as the variety of the potential users is concerned, than the former one. A strategy based on the skill prediction may consist in providing a user with the model forecast when it is expected to be good, and an alternative forecast (e.g., climatology) when it is expected to be bad. One can see in this example that the notions of "good" and "bad" forecasts are rather subjective.

The probability forecast may be expressed in a local or a regional form. An example of the the regional form can be found in Brankovic et al. (1990): the 10-day mean forecasts of 500 hPa height are clustered into three categories, using the first empirical orthogonal functions over the northern hemisphere; a probability is then assigned to each cluster. This method has the advantage of robustness, since it involves a small number of degrees of freedom: each forecast is a small set of positive numbers, the sum of which is one. It is useful for theoretical studies (it can easily take into account the bimodality of circulation patterns), but is less applicable in an operational framework. The potential users of extended range forecasts should be decision makers who have to minimize a cost function among several strategies. In this case the local formulation is more appropriate; it consists of predicting a probability distribution at a given grid point for a given field. This formulation is more difficult to verify because the model has a lot of grid points and several kind of fields. In this paper we shall restrict the analysis to the model grid points over Europe, and to the lower tropospheric temperature. Déqué (1988b) considered 3 categories, and their probabilities were simply estimated by counting the number of members of the forecast ensemble in each category. In Déqué (1988c), this probability forecast was generalized to a continuous probability distribution, and he showed that the assumption of a gaussian distribution would yield practically the same scores as the non-parametric method based on counting. However, this way of calculating the probabilities does not take into account the model imperfections, since only the ensemble spread of the forecast is available for estimating the variance of the distribution. Murphy (1990) calculated the empirical conditional probability for an observed category when a forecast category is given. Then, a probability forecast is obtained for each member of the ensemble by taking the corresponding conditional probability. Finally, the probability forecast is the average of the individual probability forecasts. This is considered as the best way to get probability forecasts, provided that the estimates for the conditional probabilities are accurate enough, which requires a large number of independent forecasts. This method cannot be easily generalized to a continuous probability distribution. In this paper, we assume that the forecast probability density is gaussian. This is generally a reasonable assumption for upper air fields and it has the advantage of requiring only two degrees of freedom (the mean and the variance) at each grid point.

In Section 2, we describe the model and the experiment. In Section 3, we present the skill of the deterministic forecasts. In Section 4, we present the probabilistic forecasts and their verification, and in Section 5, we try to optimize the skill of these forecasts by an appropriate choice of the standard deviation.

2. Description of the experiment

The forecasts we use in this paper come from a predictability experiment performed at METEO-FRANCE between 1988 and 1993. This experiment has been described in Déqué (1991b). In the present paper, we use a set of 40 forecasts (instead

of 32 in the above reference) and we study the temperature over Europe at 850 hPa (instead of the northern hemisphere 500 hPa height). Each forecast is an ensemble of 5 lagged integrations starting with ECMWF initialized analyses of day D-2 00 UTC to day D 00 UTC with a 12 hour step. The dates for day D are given in Table 1. The model is run up to day D + 44.

This model is a spectral global T42 version of the short-range forecast model operationally used at METEO-FRANCE (French weather service) during the 1984–1992 period (Coiffier et al., 1987). The package of physical parameterizations includes radiation, convection, hydrological cycle, interactive cloudiness, boundary layer physics and gravity wave drag. These parameterizations have been modified, so that this version of the model is used as a General Circulation Model (Planton et al., 1991). The vertical discretization on 20 levels is based on finite differences in hybrid coordinate. The time discretization is made with a semi-implicit scheme with a 20-min time step. The sea surface temperatures are prescribed by adding to a monthly climatology (Alexander and Mobley, 1976) a persistent anomaly, calculated with the mean day D-11/D-2 ECMWF analyses and the above climatology. Déqué (1990) showed that the extratropical skill is not significantly increased when the actual SST is used instead of the persistence of previous SST anomalies.

The skill of this model is reasonable in the short-range and slightly inferior to that of the ECMWF model (with our sample, the northern hemisphere 500 hPa heigth anomaly correlation coefficient of a 24-h forecast is 0.96 against 0.97). The mean score decreases up to day D + 10, then oscillates with a decreasing trend. There is no large difference between the score at day D + 15 and at day D + 44 (in both cases it is very low). For this reason, we average the days D + 15 to D + 44. In Déqué and Royer (1992), we have shown that the skill of such averages is very small, but statistically significant for 500 hPa height. The 30-day means corresponds more or less to the monthly means of November, December, January or February for the years 1983 to 1993. This kind of monthly mean is more representative of the long-range skill of the model than the average of days D + 1 to D + 30: as pointed out by Roads (1989), and as verified with our data, the mean D + 1 to D + 10 day model forecast is more skilful, as a predictor of the observed D + 1 to D + 30 mean, than the mean D + 1 to D + 30 day model forecast itself. The 5 lagged integrations are averaged; the model forecasts and the verifying analyses are spatially filtered to T21 truncation in spherical harmonics. As we shall see in the next section, the scores of the D + 15/D + 44 forecasts are very weak. To better illustrate the properties of the probabilistic forecast, we also include in our study the D + 5/D + 14 mean forecasts, which correspond to medium-range forecasts and are more skilful. These two verification periods will be referred to as d5/14 and d15/44 in the following.

Table 1. *Starting dates of the latest integration (i.e., day D in the text) for the 40 ensemble forecasts*

Nov.	Dec.	Jan.	Feb.
16 Oct. 1983	16 Nov. 1983	15 Dec. 1983	16 Jan. 1984
16 Oct. 1984	16 Nov. 1984	15 Dec. 1984	16 Jan. 1985
16 Oct. 1985	16 Nov. 1985	15 Dec. 1985	16 Jan. 1986
19 Oct. 1986	16 Nov. 1986	14 Dec. 1986	18 Jan. 1987
18 Oct. 1987	15 Nov. 1987	13 Dec. 1987	17 Jan. 1988
16 Oct. 1988	20 Nov. 1988	18 Dec. 1988	22 Jan. 1989
16 Oct. 1989	16 Nov. 1989	16 Dec. 1989	16 Jan. 1990
16 Oct. 1990	16 Nov. 1990	16 Dec. 1990	16 Jan. 1991
16 Oct. 1991	16 Nov. 1991	16 Dec. 1991	16 Jan. 1992
16 Oct. 1992	16 Nov. 1992	16 Dec. 1992	16 Jan. 1993

In this paper we restrict our study to a rectangular area which corresponds to Europe (65° N–40° N/10° W–35° E), and to 850 hPa temperature. This area contains 48 grid points. The advantage of this field versus the traditional 500 hPa height, is that it has a more direct impact on human activities: the monthly mean anomalies of this field, averaged over an area like France, are practically the same as the corresponding values at the surface. In the literature, the most widespread coefficient used to summarize the skill is the northern hemisphere 500 hPa height anomaly correlation; such a score is given in Déqué (1991b) for our model forecasts. Some papers, however, investigate the extended-range skill of lower troposphere temperatures (Molteni et al., 1986; Roads, 1989).

The model forecasts are compared with the ECMWF analyses of the corresponding period. Our goal is not to predict the 850 temperature field, but to perform local predictions at the different grid points of the area (in particular we do not use the strong spatial correlations of the field).

Our study could have been done for a single location, e.g., for Paris, but we use the whole European area for averaging the results and reducing the noise impact of sampling. A study of the spatial distribution of the scores over this area, based on the first 20 and the last 20 forecasts of the sample shows that the skill differences between the different grid points are essentially due to sampling.

Because of the spatial inhomogeneities of the distribution, the seasonal cycle between November and February and the presence of 10-day and 30-day averages, we have normalized our data (forecasts and verifications) with the mean and standard deviation (SD) of the ECMWF analyses based on the 10 values (one per year) of the 1983–1993 period. With this normalization procedure, the 4 calendar months are equivalent (otherwise January and February would have a larger contribution to the average); the 48 grid points are weighted by the mesh area, but the northern continental regions would have a larger contribution if the data were not normalized; the 10-day and 30-day means are directly comparable (otherwise the 10-day mean anomalies would be larger than the 30-day means). Moreover, as we shall see in Section 4, this procedure makes the deterministic forecast closer to the probabilistic one. We shall thus work with normalized values which can be considered as temperature indices: for instance, a value less than -2 indicates a very cold temperature. Though a 10-data sample is too short to test it with accuracy, it is rather reasonable to assume that the 10-day and 30-day mean temperatures have a gaussian distribution.

3. The deterministic forecast

Even at 24-h range, a meteorological forecast is inherently probabilistic since it is not always completely successful. However, the deterministic formulation is simpler to express and verify, and has been mostly used. Before performing probabilistic forecasts, we will verify that the deterministic forecast has some skill, even if very low. If the model were producing only noise (i.e., data independent of the verification data), the probabilistic forecast would be just a way to translate the absence of information into uncertainty. The time correlation coefficient gives a simple coefficient to measure the degree of statistical dependence between the forecast and the verification. Fig. 1 shows the distribution of this coefficient over the verification area for the temperature indices. For d5/14 this coefficient is minimum north of Scotland (0.2) and maximum over central Europe (0.5); the spatial average is 0.40. For d15/44, the minimum is located over Denmark (-0.1), and the maximum over Tunisia (0.5), the spatial average being 0.15. As we have mentionned in Section 2, this geographical distribution is not

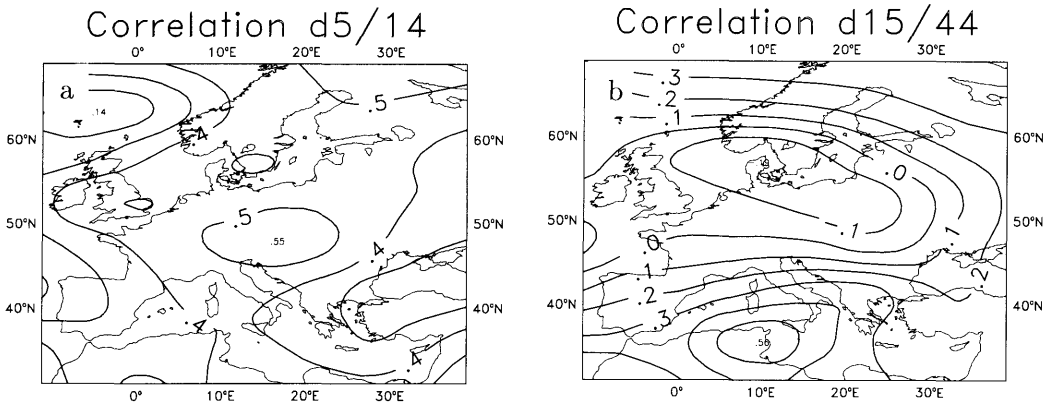

Fig. 1. Spatial distribution of the correlation coefficient between the forecast and the verification anomalies for normalized 850 hPa temperature; the forecast period is day 5–14 mean (a) and day 15–44 mean (b); contour interval 0.1.

statistically stable when we repeat the study with the first 5 or the last 5 years.

A mean correlation of 0.15 is very weak, and we need to verify that it cannot be obtained just by a favourable random sample. We have applied a procedure similar to the test applied in Déqué and Royer (1992). The 10 years of the verification data have been scrambled. Through the $10! = 3628800$ possible permutations, we have randomly selected 1000 ones. We have thus 1000 forecast/verification samples which have the same statistical properties as the original sample, except that each forecast is independent of the verification. We have then calculated the 1000 correlation coefficients and estimated a 95% confidence interval by sorting them in ascending order. For d5/14 this interval is $[-0.14, 0.15]$; for d15/44 it is $[-0.13, 0.13]$. We can thus consider that the average correlations of 0.40 and 0.15 are significantly different from zero since they are outside the confidence interval, and that the d5/14 and d15/44 forecasts contain some signal.

The correlation coefficient does not take into account the model bias. Fig. 2 gives the mean model error. The patterns are similar, but the extremes are larger for d15/44. The model is generally too cold over the eastern part, with a maximum error over Greece.

The correlation coefficient is a way to test if the forecast contains some potential information, but is not a good measure of the actual forecast skill. The mean square error (MSE) is a more accurate and less optimistic test. Murphy and Epstein (1989) proposed scaling the MSE by the MSE of the climatological forecast to get the skill score (SS):

$$\mathrm{SS} = 1 - \frac{\mathrm{MSE(for)}}{\mathrm{MSE(cli)}}. \tag{1}$$

The climatological forecast is the forecast that minimizes the MSE among all the forecasts that are independent of the verification. From a practical point of view, it is obtained by averaging the verification data of the same calendar month (to remove the influence of the annual cycle) and of the 9 other years (when performing a climatological forecast, we must ignore the verification data, otherwise we would overestimate the skill of the climatology). The MSE of the forecast can also been minimized by performing a linear regression of the model forecast by the verification forecast. Of course, this regression is calculated with the 9 other years for each case. To get a more robust formula, one can take simply:

$$\tilde{f} = cl + r(f - cl - se), \tag{2}$$

where f is the forecast, cl the climatological forecast (calculated with the 9 other years and the current calendar month), se the systematic error (calculated with the 9 other years but the 4 calendar months), and r the spatial average of the regression slope (we are not confident in its spatial pattern). The coefficient r is about 0.6 for d5/14 and 0.4 for d15/44.

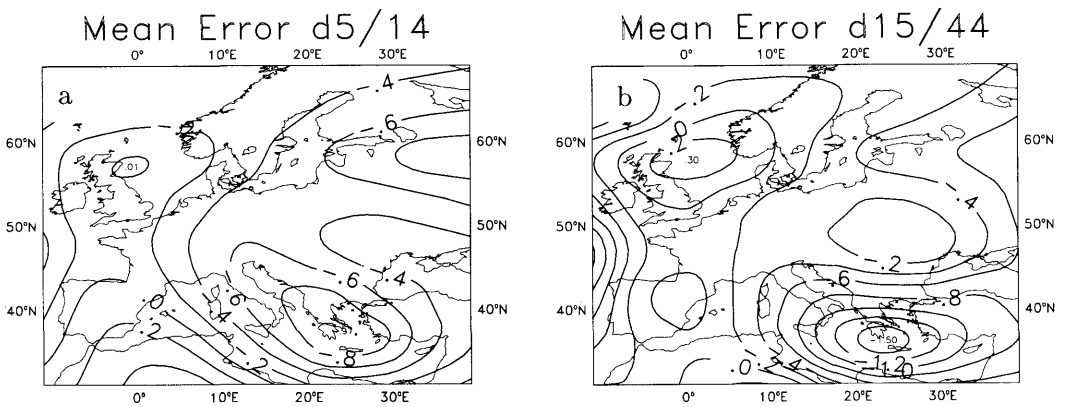

Fig. 2. As Fig. 1 for the mean forecast error; contour interval 0.2 (unit = climatological interannual standard deviation).

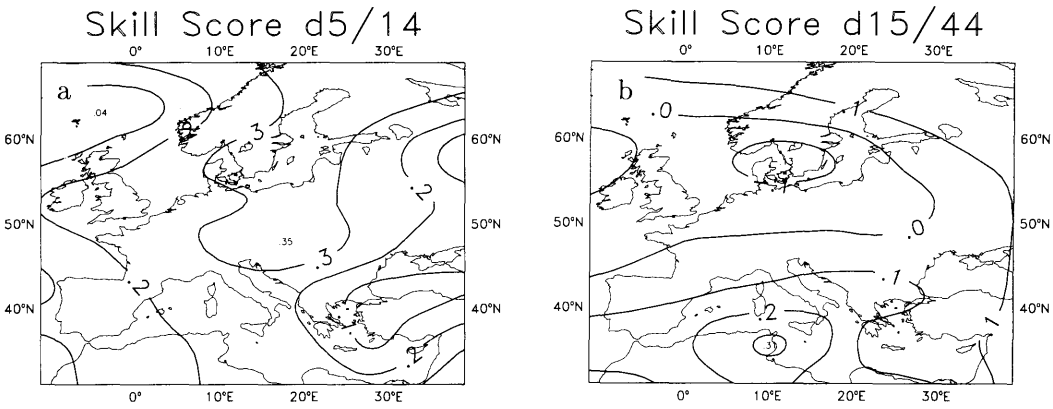

Fig. 3. As Fig. 1 for the skill score; contour interval 0.1.

The skill score of the forecasts is presented in Fig. 3. Its pattern is similar to that of the correlation coefficient. If there were no sampling error (infinite sample), the skill score should be equal to the squared correlation. The skill score has the advantage of taking into account the shortcomings due to sampling. One can see that over northern Europe the skill score of the monthly mean forecasts is negative, which means that the model forecast is poorer than the climatological forecast. A mean skill score is calculated from eq. (1), in which both MSEs are averaged over the area. It is 0.22 for d5/14 and 0.06 for d15/44. On the average, the model is thus more skilful than the climatology.

The weak but positive value of the d15/44 skill score is not steady in time: Fig. 4 shows the time fluctuations of the MSE of the model and the climatology. For d5/14, the superiority of the model is clear: the climatological MSE is less than the model MSE for only 11 cases out of the 40. For d15/44, the climatological MSE is less than the model MSE for 14 cases. Fig. 4b exhibits 4 anomalous events for which the MSE of the climatology exceeds 2. In January 1985 (case 7), a cold anomaly was observed over Europe (-1.5 on

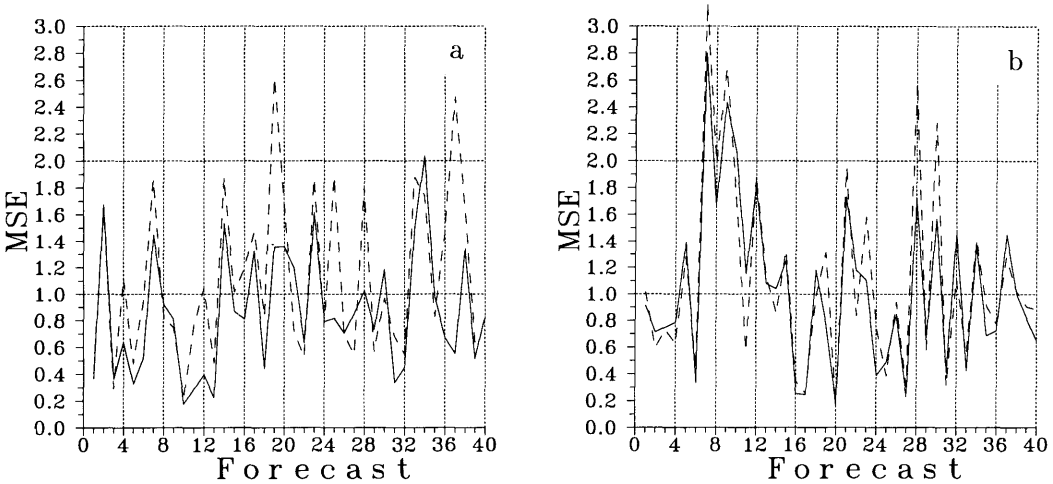

Fig. 4. Mean square errors for the 40 10-day mean (a) and 30-day mean (b) forecasts (corresponding to November 1983 through February 1993 for the 30-day means) of normalized 850 hPa temperature over Europe for the model (solid) and the climatology (dashed).

the average, minimum −2.5); the SS is 0.12. In November 1986 (case 9), a cold anomaly was located in the north-western part of the area (minimum −2.5); the SS is 0.05. In February 1990 (case 28), a warm anomaly was observed (1.5 on the average, maximum 2); the SS is 0.33. In December 1990 (case 30), a cold anomaly was located over the western part of the area (minimum −2.5); the SS is 0.31. The model is relatively successful for these extreme events. The SS varies between −1.03 for January 1986 (case 11) because of a small value of the climatology MSE, and 0.49 for February 1989 (case 24) which is associated with small values of both MSEs.

4. The probabilistic forecasts

4.1. The ranked probability score

The smallness and variability of the skill naturally lead to the expression of the forecast as a probability distribution, i.e., as a set of values $(p_1, ..., p_n)$ which are the probabilities for the events $1, ..., n$. The observation is the vector $(o_1, ..., o_n)$, where $o_i = 1$ when the event i occurred, $o_i = 0$ otherwise. The best probability forecast is the closest to the observation. A natural score is the squared distance between the vectors $(p_1, ..., p_n)$ and $(o_1, ..., o_n)$. However, when the events correspond to ranked categories, the ranked probability score (RPS), introduced into meteorology by Epstein (1969), takes into account the fact that an error in predicting a category close to the observed one is less important than an error in predicting a category far from the observed one. This score is simply the squared distance between the cumulated vectors $(P_1, ..., P_n)$ and $(O_1, ..., O_n)$, where $P_i = \sum_{j \leqslant i} p_j$ and $O_i = \sum_{j \leqslant i} o_j$. Following Murphy and Daan (1985), we have slightly modified Epstein's original definition by setting:

$$\mathrm{RPS} = \frac{1}{n-1} \sum_{i=1}^{n} (P_i - O_i)^2. \tag{3}$$

The division by $n-1$ provides a value less than 1 whatever n (the sum involves $n-1$ terms since the term $i = n$ is zero) and allows us to consider the limit of the RPS as n tends to infinity (continuous probability forecast). Moreover, in the next subsection, we use a skill score based on the ratio of two RPSs: the division by $n-1$ does not modify such a score.

If the forecast is a gaussian distribution of mean x_f and SD σ_f, if the observation is x_o, and if the n categories are defined by the $n-1$ thresholds $s_1 < \cdots < s_{n-1}$, then the RPS is:

$$\mathrm{RPS} = \frac{1}{n-1} \sum_{i=1}^{n-1} [\Phi((s_i - x_f)/\sigma_f) - \theta(s_i - x_o)]^2 \tag{4}$$

where Φ is the cumulative density of a standard gaussian distribution (also called erfc function), and θ the step function (also called Heaviside function) defined by $\theta(x) = 0$ when $x \leqslant 0$ and $\theta(x) = 1$ when $x > 0$.

The thresholds s_i are chosen here so that the n categories are equiprobable with respect to the climatology, i.e., the s_i are the n-quantiles of the climatological distribution. If we assume that the climatology distribution of the parameter (here the normalized 850 hPa temperature at a given grid point) is a gaussian distribution of mean x_c and SD σ_c, the thresholds are:

$$s_i = x_c + \sigma_c \Phi^{-1}(i/n). \tag{5}$$

Thus, the RPS of a gaussian forecast is a function of n and of 5 parameters x_f, σ_f, x_o, x_c, and σ_c. In fact, these parameters can be reduced to the 3 normalized parameters:

$$u_f = \Phi((x_f - x_c)/\sigma_c),$$

$$u_o = \Phi((x_o - x_c)/\sigma_c),$$

$$t = \frac{2}{\pi} \tan^{-1}(\sigma_f/\sigma_c).$$

u_f is the normalized forecast mean, u_o the normalized observation, and t the normalized forecast SD ($t = 0$ corresponds to a purely deterministic forecast and $t = 1$ corresponds to a uniform probability forecast). These 3 parameters range in $[0, 1]$. The choice of $\tan^{-1}$ is quite arbitrary and does not simplify the formulae: it only allows us to map the ratio σ_f/σ_c into the same finite interval $[0, 1]$. With such notations, eq. (4) becomes:

$$\mathrm{RPS}(n, u_f, u_o, t)$$

$$= \frac{1}{n-1} \sum_{i=1}^{n-1} \left[\Phi\left(\frac{\Phi^{-1}(i/n) - \Phi^{-1}(u_f)}{\tan(\pi t/2)} \right) - \theta(i/n - u_o) \right]^2. \quad (6)$$

Grid point values of 850 hPa temperature can be used to produce categorical forecasts, simply by predicting the category to which the value belongs, instead of the value itself. When n equals to 2, the categorical forecast consists of predicting the sign of the anomaly. The deterministic forecasts studied in Section 3 can be considered as the extreme case of an infinite number of categories: we predict the ratio of the anomaly by the climatological SD, which is a kind of label of the category to which the forecast value belongs. From eq. (6), we deduce the RPS of the categorical case:

$$\mathrm{RPS}(n, u_f, u_o, 0) = \frac{1}{n-1} \sum_{i=1}^{n-1} [\theta(i/n - u_f) - \theta(i/n - u_o)]^2. \quad (7)$$

When $n = 2$, the RPS is 1 when u_f and u_o belong to the same category, and is 0 otherwise. When n tends to infinity, the RPS of the deterministic forecast converges to:

$$\mathrm{RPS}(\infty, u_f, u_o, 0) = |u_f - u_o| \quad (8)$$

instead of the mean square error (or L_2 norm), we get the mean absolute error (MAE) which has similar properties (L_1 norm).

4.2. The probability skill score

The RPS we have introduced is thus a measure of the distance between two probability forecasts. But, as for the MSE, we need a reference value in order to translate it into a relative measure of the forecast quality. The mean RPS for a large number of cases is obtained by taking the expectation of eq. (6) relative to the distribution of (u_o, u_f, t). If we assume that the forecasts are independent of the verifications, we can calculate the expectation with respect to u_o separately. Assuming that the observation x_o has a gaussian distribution, one can demonstrate that, among the forecasts without any skill (i.e., for which u_f and t are independent of the verification u_o), the forecast which minimizes the mean RPS is given by $u_f = \frac{1}{2}$ and $t = \frac{1}{2}$. If we come back to the original dimensional values, this means $x_f = x_c$ and $\sigma_f = \sigma_c$, and such a forecast is simply the climatological probability forecast. Its RPS is:

$$\mathrm{RPS}_c(n, u_o) = \mathrm{RPS}(n, \tfrac{1}{2}, u_o, \tfrac{1}{2}) = \frac{6I^2 - I(n-1) + (n-1)(2n-1)}{6n(n-1)}, \quad (9)$$

where I is the largest integer less than or equal to nu_o. Similarly to the skill score introduced by eq. (1), we now define the probability skill score (PSS):

$$\mathrm{PSS} = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS}_c}. \quad (10)$$

With the same convention as for the SS, when we aggregate different grid points or different forecasts, the average is done separately for both RPSs. One can see that this definition is insensitive to a constant factor applied to the RPS. This skill score has been introduced by Murphy and Daan (1985) as I_{RPS}.

Averaging eq. (9) yields the mean RPS_c: $(n+1)/6n$. This implies that the RPS tends to decrease as n increases; this behaviour is due to our normalization by $n-1$; with Epstein's original definition the RPS increases as n increases. It is natural to consider that a 500-category forecast is superior to a 2-category forecast, since one can always deduce the latter from the former but not vice-versa. However, the PSS of the former is not necessarily greater than the PSS of the later, because its RPS is divided by a smaller value. If we use the PSS to compare 2 forecasts with different n, one must keep in mind that we do not compare their potential utility, but the information increase with respect to the best trivial forecast of the same type.

The same discussion as above holds for the categorical forecast. If we compare a probability forecast with the categorical forecast obtained with the mean of the probability distribution, the RPS of the latter is generally larger. Indeed, the RPS has been introduced to favour the forecasts which express their uncertainty. Moreover, one can stress, as above, that a probability forecast can

always be transformed into a categorical one. But one can also look for the trivial categorical forecast which minimizes the RPS. The same kind of calculations as above leads to $u_f = \frac{1}{2}$ and $t = 0$ which is the climatological deterministic forecast introduced in Section 3. Its mean RPS is $n/4(n-1)$ when n is even, and $(n+1)/4n$ when n is odd; in the latter case, the central category exists and the climatology forecast is more successful. This mean RPS is, as expected, greater than the mean RPS of the climatological probability forecast, $(n+1)/6n$. We introduce the categorical skill score (CSS) in the same way as in eq. (10), taking the ratio of the RPS of the categorical forecast by the RPS of the climatological categorical forecast.

4.3. *Several kinds of probability forecasts*

In contrast to the case of the categorical or probabilistic climatological forecast, the RPS of the model forecast given by eq. (6) cannot be given by a simple analytical formula. We have discretized the interval [0, 1] into 53 values for u_f and u_o: the choice of a prime number has been done to avoid the discontinuity of the step function θ. This discontinuity has no importance for integral quantities, but produces small alterations when the integrals are discretized: e.g., we can write $\theta(u) = 1 - \theta(-u)$ except for $u = 0$. For t, the interval has been discretized in 100 values since the minimization of Section 5 will require a better precision. The RPS is calculated and tabulated. With our data, the computations based on this tabulation and a linear interpolation lead to an error of less than 1%. We have tabulated the RPS for $n = 2$, the lowest possible value, $n = 3$, $n = 4$, $n = 5$, because terces and quints are often used in categorical forecasts, and $n = 500$. With this last value, the difference with the limit when $n = \infty$ is estimated to be less than 1%; we can consider this case as a suitable approximation of the continuous probability forecast.

Table 2. *Skill score (CSS for the first row, PSS for the others) of different kinds of forecast (categorical and probabilistic) as a function of the number of categories n for d5/14 model forecasts*

n	2	3	4	5	500
categorical	0.33	0.13	0.20	0.16	0.16
prob. (model spread)	0.05	0.01	0.02	0.02	0.02
prob. (mean model spread)	0.07	0.03	0.04	0.04	0.04
prob. (climatological SD)	0.16	0.12	0.12	0.12	0.11
prob. ("improved")	0.16	0.14	0.13	0.13	0.13

Table 2 (for d5/14) and Table 3 (for d15/44) give the skill scores (CSS or PSS) as a function of n for different kinds of probability forecasts. The first row corresponds to the categorical forecast calculated from eq. (2) (Section 3); in fact this is a statistical adaptation of the model deterministic forecast. The oscillation between the odd and even values of n is due to the fact that for n even, there is no central category. The CSS for $n = 500$ are slightly less than the SS given in Section 3 (0.22 and 0.06 respectively). These differences are explained by the fact that the SS is based on the MSE and the CSS on the MAE. The second row is for a probability forecast based on the mean ensemble forecast (after correction of the systematic error) and of the ensemble SD based on the 5 individual integrations for each forecast. The PSS is very poor. In order to check if the ensemble spread contributes to improve the forecast, we have replaced the case-by-case model spread by the mean spread calculated with the 40 forecasts. The results (third row) are somewhat improved: this shows that the variability of the model spread does not bring any useful information, or, at least, that the sampling noise contained in it degrades the probability forecast. This is not surprising since a sample of only 5 values is much too short to give a reliable estimate of the variance. The absence of

Table 3. *As Table 2, but for d15/44 model forecasts*

n	2	3	4	5	500
categorical	0.21	0.04	0.10	0.03	0.04
prob. (model spread)	−0.14	−0.13	−0.13	−0.14	−0.13
prob. (mean model spread)	−0.12	−0.09	−0.10	−0.10	−0.09
prob. (climatological SD)	−0.02	−0.02	−0.02	−0.03	−0.03
prob. ("improved")	0.03	0.03	0.02	0.02	0.01

correlation between the spread and the skill in extended-range forecasts has been shown in Déqué (1991b). The use of the model spread in probability forecast is based on the idea that the model error is mostly due to the amplification of the small initial uncertainties. However, since the model has some deficiencies, its ensemble spread underestimates the total error. We give in the fourth row the PSS of the model forecast associated with the climatological SD. The scores are improved, but the model skill scores remains below that of climatology for d15/44.

We have tried to optimize our probabilistic forecasts, as we have done for the deterministic ones in Section 3. Let us come back to the general formalism of a n-category forecast. We suppose that the forecast X_f and its verification X_o are dependent random variables. We look for a set of n formulae $p_1(x_f), ..., p_n(x_f)$, which give the probabilities as a function of the ensemble mean model forecast x_f (we do not take into account the ensemble spread). We would like to choose these functions in order to minimize the expectation of the RPS. Some straightforward probabilistic calculations show that this is obtained when the $p_j(x_f)$ are the conditional probabilities of the observed category j when the model forecast is x_f (i.e., prob$(o_j = 1 \mid X_f = x_f)$).

With a short sample of forecasts and verification, it is impossible to estimate with robustness the distribution of the verification data for each value of the forecast. A first method to obtain stable results is to use a small number of categories for x_f, or, equivalently, for u_f. The simplest case is $n = 2$ and two categories for u_f. We have estimated the mean RPS from the 2x2 contingency table of (u_o, u_f). For d5/14, the PSS is 0.14, and for d15/44, it is 0.03. To avoid the overestimation of the PSS, one should estimate the conditional probabilities with the 9 other years for each forecast: this procedure reduces the estimates of the PSS (0.10 for d5/14 and 0.01 for d15/44). The use of empirical contingency tables does not improve the skill for d5/44 and brings a small and not robust improvement for d15/44. One can easily guess that with more than two categories, the problems of robustness would be even more acute.

Another method can be based on the assumption that the pair forecast-verification (X_f, X_o) has a gaussian distribution. In this case, we know that the conditional distribution of the verification X_o when the forecast X_f equals x_f is also a gaussian distribution; its mean is the regression of the verification by the forecast (given by eq. (2)) and its SD is the SD of the verification (i.e., the climatological SD) multiplied by $\sqrt{1-\rho^2}$ where ρ is the correlation between forecast and verification. For the sake of robustness, we have taken the spatial RMS of ρ. To avoid the overestimation of the PSS, one should estimate ρ with the 9 other years for each forecast; it ranges between 0.36 and 0.42 for d5/14, and 0.22 and 0.26 for d15/44; to simplify the procedure we have taken for ρ, 0.4 and 0.2, respectively. The results of this "improved" probability forecast are given in the 5th row of Tables 2 and 3. It gives only a small improvement with respect to the method based on the mean model forecast and the climatological SD.

There is a slight skill decrease as the number of categories n increases (with an odd/even oscillation for the CSS). The PSS of the improved probabilistic forecast is slightly less than the CSS. We cannot compare the PSS with the CSS with accuracy (nor are the SS and the CSS fully comparable), but we can emphasize that no additional information is brought in the probabilistic case versus the categorical one (the mean of the distribution is the same and the SD is climatological). The model RPS is improved by the use of a probabilistic formulation, but the climatology RPS is improved to a larger extent. We have checked that this kind of behaviour (values of the CSS and PSS, decrease with n, oscillation of the CSS) can be reproduced by generating a random series of pseudo forecast/verification pairs (i.e., gaussian pairs having the same mean, variance, and correlation as the original series).

5. Optimisation of the standard deviation

We have seen in Tables 2 and 3 that the use of the SD of the observations to represent the uncertainty of the model forecast provides a better score than the use of the model spread. The score may be slightly improved when this climatological SD is multiplied by a constant factor less than 1. The use of a smaller SD than the climatological one is consistent with the use of an expectation departing from the climatological mean. In this kind of forecast, the mean of the distribution varies from one case to another, but the SD remains constant.

In this section, we shall look for the possibility of forecasting a SD which should lead to a better score than the improved probability forecast described in Section 4.

Let us consider the best choice we could do when the number of categories n is fixed. For a given value of the forecast and verification (u_f and u_o) we choose the value of t which minimizes the RPS given by eq. (6). Let us name it $t_m(u_f, u_o)$. This choice is optimal, but academic: it cannot be performed for actual forecasts since one ignores the verification value u_o; on the other hand, if we know u_o, the best forecast is given by $u_f = u_o$ and $t = 0$. However this choice gives the upper boundary of the PSS among all the possibilities for the SD, when the mean of the distribution if u_f (u_f is the normalized value of the result of the regression given by eq. (2)).

With two categories ($n = 2$), the calculation of $t_m(u_f, u_o)$ is simple:

- if ($u_f \leqslant 0.5$ and $u_o \leqslant 0.5$) or ($u_f > 0.5$ and $u_o > 0.5$), the forecast is good, then $t_m(u_f, u_o) = 0$, and the corresponding minimum RPS is 0 (PSS = 1);
- if ($u_f \leqslant 0.5$ and $u_o > 0.5$) or ($u_f > 0.5$ and $u_o \leqslant 0.5$), the forecast is bad, then $t_m(u_f, u_o) = 1$, and the corresponding minimum RPS is 0.25 (PSS = −0.5).

On the average, the PSS of such an academical forecast is $1 - 1.5 f_{bad}$ where f_{bad} is the frequency of bad forecast.

As n increases, the expression of $t_m(u_f, u_o)$ and of the corresponding minimum RPS becomes more complicated. When n tends to infinity, $t_m(u_f, u_o)$ tends to a continuous function. Fig. 5 shows the distribution of $t_m(u_f, u_o)$ and of the corresponding PSS for $n = 500$. If we change u_f and u_o to $1 - u_f$ and $1 - u_o$, the RPS is unchanged. Thus, the 2 patterns are symmetrical with respect to the point ($u_f = 0.5$, $u_o = 0.5$). They seem symmetrical with respect to the diagonals, but the symmetry is not exact. On the diagonal $u_f = u_o$, we have obviously $t_m = 0$ and PSS = 1. Roughly, t_m and the PSS depend on $|u_f - u_o|$. For $|u_f - u_o|$ large enough, we have $t_m = 1$ and the PSS is −0.5. In fact the value of σ_f corresponding to $t_m(u_f, u_o)$ can be interpreted as a distance between x_f and x_o.

The PSS of our 40 forecasts has been calculated with this optimal SD. The results are displayed in Table 4. One can see that when the model forecasts are associated with an appropriate SD, the skill is widely enhanced. This enhancement is more pronounced with a small number of categories.

We have tried to predict this optimal SD in the case of 500 categories. This kind of prediction is equivalent to the forecast skill prediction (e.g., Palmer and Tibaldi, 1988) within the frame-

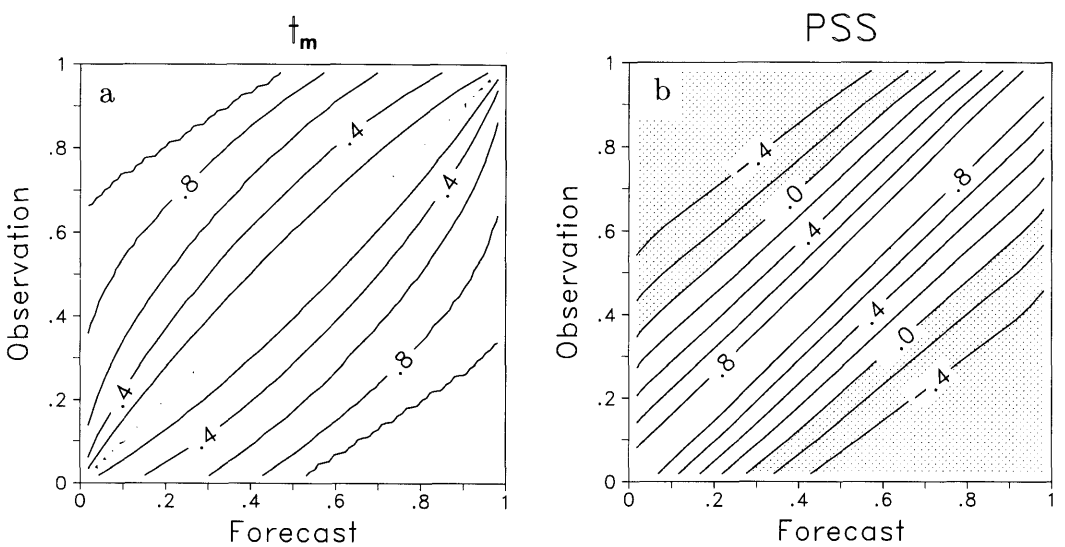

Fig. 5. Distribution of the optimum standard deviation t_m (a), and of the corresponding maximum PSS (b), as a function of the forecast u_f and observation u_o. Contour interval 0.2, the negative values are shaded.

Table 4. *Maximum probability skill score one can get when using the model forecast as the mean of the probability distribution and when varying the standard deviation; n is the number of categories*

n	2	3	4	5	500
d5/14	0.64	0.44	0.39	0.35	0.28
d15/44	0.53	0.32	0.27	0.24	0.18

work of probability forecasts. Indeed, t_m is well correlated with the absolute error (we found a correlation of 0.88 for d5/14 and of 0.82 for d15/44). Unfortunately, the absolute error is unknown at the time of the forecast. We have tried to correlate t_m with different classical predictors of the forecast skill (Déqué, 1991b). The spread of the forecast (measured by the ensemble SD) is not correlated at all (0.03 and -0.04 for d5/14 and d15/44, respectively). The amplitude of the forecast (measured by the absolute forecast anomaly) is negatively correlated (-0.15 and -0.06, respectively) as expected, but these correlations are not significant.

The absence of correlation between t_m and the spread can be explained by the fact that the ensemble spread at a given grid point is too poorly estimated with only 5 values. We have used more stable indices of the forecast spread. In Stroe and Royer (1993), the growth of the global MSE of the 500 hPa height is modelled with only two parameters, representing a saturation value and a saturation rate of the error growth curve. We have thus 2 indices for each of the 40 forecasts, which indicate how stable the initial situation is. The correlation between t_m and each of the two indices is however not significant.

Large-scale features of the initial situation, as shown in the study of Palmer (1988) on the sign of the PNA, or as shown by composites of skilful forecasts (Déqué, 1991b), may be connected to the forecast skill. We have carried out an empirical orthogonal function (EOF) analysis of the northern hemisphere 500 hPa height. The data are the 40, 10-day mean ECMWF analyses corresponding to the 10 days before the starting date of each forecast. The first 3 EOFs explain 17%, 14%, and 10% of the total variance. These EOFs have been used as predictors in a linear regression of t_m. The squared multiple correlation is 0.08 for d5/14 and 0.07 for d15/44.

These statistical predictions have been verified with a cross-validation procedure: case-by-case predictions of t_m are performed, using the remaining 39 cases as a training sample; the mean square difference between the predicted and the actual value is compared with the variance of the predictand. This procedure has been applied to test the correlation coefficients (simple regression) and the squared multiple correlation coefficients (multiple regression). In all our attempts, we have no skill at all, i.e., the MSE is greater than the variance. A similar study based on the first 10 EOFs is no more successful.

Thus, no attempt of predicting the optimal SD has lead to a significant result. This result is not very surprising. The optimal SD is strongly correlated with the absolute forecast error. If we were able to predict the optimal SD at a given grid point and for a given forecast, we would be able to predict the forecast error. This error is unpredictable, unless we have a more skilful model to perform a forecast closer to the verification. The studies on the prediction of the skill do not attempt to predict the error map, but an index of the spatial average of the forecast quality (anomaly correlation coefficient or MSE). We can however try to predict a uniform SD on our domain, without trying to predict its spatial distribution. With a larger sample, we could try to locate the area with maximum predictability, and thus refine the spatial pattern of the SD.

In Sections 3 and 4, we have used a uniform value of the correlation coefficient ρ or of the regression slope r (in eq. (2)), since we have considered that the spatial fluctuations of these parameters were essentially due to sampling. Accordingly, we have performed a posteriori optimal probability forecasts in which the parameter t (normalized SD) is uniform over Europe (but varies among the 40 cases). Thus, for each forecast, the spatial averaged RPS is calculated as a function of t, and the value t_{opt} which corresponds to the minimum RPS is stored. It is interesting to note that t_{opt} is very close to the spatial average of t_m: the correlation between the two coefficients is 0.98 for d5/14 and d15/44. When performing probability forecasts with this coefficient, we get a PSS of 0.15 for d5/14 and 0.04 for d15/44. These values are not much greater than those obtained in Section 4. The values for t_{opt} need to be predicted with accuracy (they are unknown a priori) to expect

any improvement. We have tried the same predictors as for t_m, but without any large success. The regressions which provide a correlation greater than 0.5 were not robust in cross-validation, and the robust regressions we used provide PSSs less than the values found in Section 4 (0.13 for d5/14 and 0.01 for d15/44).

6. Summary and perspectives

We have studied 10-day and 30-day forecasts of an index of the lower tropospheric temperatures. These forecasts are local, although the scores are averaged over Europe at the end of the process. These forecasts are not independent of the verifying analyses and, after a suitable processing which minimizes the MSE (subtraction of the systematic error and reduction of the anomaly amplitude), the SS is positive but very small for the monthly means. We have performed a local probability forecast based on a gaussian distribution with the deterministic forecast as a mean, and various possibilities for the SD. A particular case is the categorical forecast for which the SD is zero. The skill of the forecasts is measured with the RPS. We have shown that, in the absence of any skill, the RPS is minimum with the climatology forecast. We have thus applied two skill scores for the full probabilistic and for the categorical forecast. The use of the model spread leads to a smaller PSS than the use of the climatological SD. We have shown that the PSS is maximum when we use, for the probability forecast, the conditional probability of the verification. Under the gaussian hypothesis, this method leads to practically the same results and the same PSS as the forecast with the climatological SD. Finally, we have shown that the PSS could be widely increased by taking and adequate SD (and keeping the same mean for the distribution). Unfortunately our attempts to obtain a greater PSS than with the conditional probabilities, by predicting an optimal SD, were not successful.

We are still far from the operational use of probability extended range forecasts, but our study provides some useful statements. The PSS is not very sensitive to the number of categories used in the formulation, and thus there is only a small gain in reducing the number of categories. Providing a user with a large number of categories allows him to gather the categories (and add the corresponding probabilities) according to his needs. The categorical forecast is worse than the probabilistic one in terms of RPS, but, since the categorical climatological forecast is also worse than the probabilistic categorical one, we get the opposite result when we scale the error by the climatology (i.e., PSS < CSS). We consider however that the best formulation is the probabilistic one, since the categorical forecast is simply recovered from the probabilities by selecting the most probable category.

We are faced to the same problem as the prediction of the forecast skill. If we were able to predict the SD (rather than taking a climatological value), the probability forecast would be useful, even if the deterministic one were poor. There are at least 3 possible ways of improving the probability forecasts. The 1st way is to use empirical conditional probabilities. This requires larger samples than ours, but can be applied to short- and medium-range forecasts (there are hundreds of winter forecasts available in the ECMWF archive). The 2nd way is to use a better estimate of the model spread as a predictor of the skill (and thus of the SD). This requires ensemble forecasts of more than 20 individual model integrations, and would make the experiment very expensive, since we need also a sufficient number of independent forecasts. The 3rd way is to get rid of the local gaussian formulation, and to try to forecast large scale patterns (e.g., clusters or EOFs). The weak but significant skill-spread correlations obtained when the spread is measured by the northern hemisphere forecast agreement (Murphy, 1990) in the medium range, and the results of northern hemisphere clustering (Brankovic et al., 1990), are somewhat encouraging.

7. Acknowledgements

The authors are particularly indebted to all their colleagues from METEO-FRANCE who developed the numerical forecast model, and to ECMWF for providing global daily analyses. Thanks are also due to M. Chipperfield for his help in wording. This work was partly supported by the Commission of European Communities (contract EV5V-CT92-0121), by PNEDC (French National Program for the Study of Climate Dynamics), and by the Regional Council of Midi-Pyrénées.

REFERENCES

Alexander, R. C. and Mobley, R. L. 1976. Monthly average sea surface temperatures and ice-pack limits on a 1° global grid. *Mon. Wea. Rev.* **107**, 896–910.

Brankovic, C., Palmer, T. N., Molteni, F., Tibaldi, S. and Cubasch, U. 1990. Extended-range predictions with ECMWF models: time lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.* **116**, 867–912.

Coiffier, J., Ernie, Y., Geleyn, J. F., Clochard, J., Hoffman, J. and Dupont, F. 1987. The operational hemispheric model at the french meteorological service. *J. Meteor. Soc. Japan*, special NWP symposium volume, 337–345.

Déqué, M. 1988. 10 day predictability of the northern hemisphere winter 500 mb height by the ECMWF operational model. *Tellus* **40A**, 26–36.

Déqué, M. 1991. 44-day ensemble forecasts with the T42-L20 french spectral model. In: *Proceedings of ECMWF workshop on new developments in predictability*, 13–15 November 1991, ECMWF, Shinfield Park, Reading, 247–263.

Déqué, M. 1990. Impact of prescribed sea surface temperatures on extended range forecasting. *J. Marine Systems* **1**, 61–70.

Déqué, M. 1988. The probabilistic formulation: a way to deal with ensemble forecasts. *Annales Geophysicae* **6**, 217–224.

Déqué, M. 1988. Probabilistic monthly mean predictions using forecast ensembles. In: *Proceedings of ECMWF workshop on predictability in the medium and extended range*, 16–18 May 1988, ECMWF, Shinfield Park, Reading, 119–134.

Déqué, M. 1991. Removing the model systematic error in extended range forecasting. *Annales Geophysicae* **9**, 242–251.

Déqué, M. and Royer, J. F. 1992. The skill of extended-range extratropical winter dynamical forecasts. *J. Climate* **5**, 1346–1356.

Epstein, E. S. 1969. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.* **6**, 762–769.

Hoffman, N. R. and Kalnay, E. 1983. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus* **35A**, 100–118.

Kalnay, E. and Dalcher, A. 1987. Forecasting the forecast skill. *Mon. Wea. Rev.* **115**, 349–356.

Leith, C. E. 1974. Theoretical skill of monte carlo forecasts. *Mon. Wea. Rev.* **102**, 409–418.

Milton, S. F. and Richardson, D. S. 1991. 30-day dynamical forecasts at the UK Meteorological Office. In: *Proceedings of ECMWF workshop on new developments in predictability*, 13–15 November 1991, ECMWF, Shinfield Park, Reading, 205–245.

Molteni, F., Cubasch, U. and Tibaldi, S. 1986. 30- and 60-day forecast experiments with the ECMWF spectral models. In: *Proceedings of ECMWF workshop on predictability in the medium and extended range*, 17–19 March 1986, ECMWF, Shinfield Park, Reading, 51–107.

Murphy, A. H. and Daan, H. 1985. Forecast evaluation. In: *Probability, statistics and decision making in the atmospheric sciences*, edited by A. H. Murphy and R. W. Katz, Westview Press, London, 379–438.

Murphy, A. H. and Epstein, E. S. 1989. Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.* **117**, 572–581.

Murphy, J. M. 1990. Assessment of the practical utility of extended range ensemble forecasts. *Quart. J. Roy. Meteor. Soc.* **116**, 89–125.

Palmer, T. N. 1988. Medium and extended range predictability and stability of the Pacific/North American mode. *Quart. J. Roy. Meteor. Soc.* **114**, 691–713.

Palmer, T. N., Brankovic, C., Molteni, F. and Tibaldi, S. 1990. Extended range predictions with ECMWF models. I Interannual variability in operational model integrations. *Quart. J. Roy. Meteor. Soc.* **116**, 799–834.

Palmer, T. N. and Tibaldi, S. 1988. On the prediction of forecast skill. *Mon. Wea. Rev.* **116**, 2453–2480.

Planton, S., Déqué, M. and Bellevaux, C. 1991. Validation of an annual cycle experiment with a T42-L20 GCM. *Climate Dyn.* **5**, 189–200.

Roads, J. O. 1989. Dynamical extended range forecasts of the lower tropospheric thickness. *Mon. Wea. Rev.* **117**, 3–28.

Shea, D. J. and Madden, R. A. 1990. Potential for long-range prediction of monthly mean surface temperatures over North America. *J. Climate* **3**, 1444–1451.

Sirutis, J. and Miyakoda, K. 1990. Subgrid scale physics in 1-month forecasts. Part I: experiments with four parameterization packages. *Mon. Wea. Rev.* **118**, 1043–1064.

Stroe, R. and Royer, J. F. 1993. Comparison of different error growth formulas and predictability estimation in numerical extended-range forecasts. *Annales Geophysicae* **11**, 296–316.

Tracton, M. S., Mo, K., Chen, W., Kalnay, E., Kistler, R. and White, G. 1989. Dynamical Extended Range Forecast (DERF) at the National Meteorological Center. *Mon. Wea. Rev.* **117**, 1604–1635.

Yamada, S., Maeda, S., Kudo, T., Iwasaki, T. and Tsuyuki, T. 1991. Dynamical one-month forecast experiments with the JMA global prediction model. *J. Meteor. Soc. Japan* **69**, 153–159.