# 44-DAY ENSEMBLE FORECASTS WITH THE T42-L20 FRENCH SPECTRAL MODEL

Michel Déqué
METEO-FRANCE / CNRM
Toulouse, France

**Abstract**

A set of 32 winter forecasts has been produced with a frozen version of the French spectral model. Each forecast consists of 5 lagged integrations. The impact of the ensemble technique, the systematic error correction, and the choice of the reference climatology are discussed. The 500 hPa anomaly correlation over the northern hemisphere is significantly positive, but weak on the average. Possible predictors of this skill variability are investigated. The skill is also measured when the forecasts are given in probability form.

## 1. INTRODUCTION

Most recent numerical weather prediction models have been integrated beyond the limit of useful forecasts and have shown that some skill exists beyond the medium range (*Tracton et al* 1989, *Murphy* 1990, *Sirutis and Miyakoda* 1990, *Palmer et al.*). However this average skill results of a mixing of particularly good and very bad forecasts. In order to produce statistically significant results, one needs large samples of independent forecasts. In *Déqué* (1988a,1988b), a set of 5 forecasts was used to measure the impact of resolution increase (T21 to T42), and the skill of probability forecasts. A larger set of 21 forecasts has been used in *Déqué* (1991) to measure the impact of the correction of the systematic error : poor estimates of the model bias have a negative impact on the scores of the corrected forecasts, and a sufficient number of homogeneous and independent forecasts is required to apply such a correction. In this presentation, we have extended the previous set to 32 forecasts, in order to better capture the mean skill (section 3), and to study the skill variability (section 4).

## 2. DESCRIPTION OF THE EXPERIMENT

The predictability experiment consists of a series of 160 extended range 46-, 45-, or 44-day integrations with a global T42 climate version of the model of the French weather service "Emeraude" (*Coiffier et al.*, 1987). This spectral model uses a comprehensive set of physical parameterizations including radiative and hydrological cycles, interactive cloudiness, boundary layer physics, convection (*Bougeault*, 1985), and gravity wave drag. The vertical discretization

| | | | |
|---|---|---|---|
| 1983/10/16 | 1983/11/16 | 1983/12/15 | 1984/01/16 |
| 1984/10/16 | 1984/11/16 | 1984/12/15 | 1985/01/16 |
| 1985/10/16 | 1985/11/16 | 1985/12/15 | 1986/01/19 |
| 1986/10/19 | 1986/11/16 | 1986/12/14 | 1987/01/18 |
| 1987/10/18 | 1987/11/15 | 1987/12/13 | 1988/01/17 |
| 1988/10/16 | 1988/11/20 | 1988/12/18 | 1989/01/22 |
| 1989/10/16 | 1989/11/16 | 1989/12/16 | 1990/01/16 |
| 1990/10/16 | 1990/11/16 | 1990/12/16 | 1991/01/16 |

Table 1: Starting dates of the latest integration of each ensemble for the 32 forecasts.

is based on finite differences in hybrid coordinate (*Simmons and Burridge*, 1981) with 20 levels up to 1 hPa, and the time discretization on a semi-implicit scheme (20 min. time step). The annual climatology of this model has been described by *Planton et al.* (1991). The daily sea surface temperatures (SST) are prescribed by a linear interpolation between two climatological monthly means (*Alexander and Mobley*, 1976) to which is added a constant SST anomaly, calculated by averaging the daily SST ECMWF analyses for the 10 days preceding the beginning of the integration. A hindcast experiment (*Déqué*, 1990) has shown that the extratropical skill of this model is not significantly increased by the use of daily observed SST as a boundary condition.

Each forecast is an ensemble of 5 integrations of the model obtained by the lagged average technique (*Hoffman and Kalnay*, 1983). The initial conditions are the ECMWF initialized analyses of day −2 00 Z, day −2 12 Z, day −1 00 Z, day −1 12 Z, and day 0 00 Z. The dates for day 0 are given in Table 1. They correspond to the middle of October, November, December or January. An ensemble mean forecast consists of the average of the 5 results of the lagged integrations. The 5 integrations are run until day 44. The period day 15/day 44 corresponds approximately to a calendar month, and an ensemble forecast will be referred to by the name of this month : e.g. the integrations starting between 14 and 16 October 1983 will constitute the November 1983 ensemble forecast. The fields are expanded on a global 64 x 32 longitude latitude Gaussian grid after a reduction of the spectral truncation to T21.

## 3. MEAN SKILL

### 3.1 Verification score

The model forecasts are compared with the corresponding ECMWF initialized analyses expanded on the same grid (in T21 truncation as for the forecasts). The forecast skill is measured

with the spatial anomaly correlation coefficient (ACC) over the northern hemisphere (in fact north of $20°N$). The anomaly is obtained by subtracting a climatology for the current month based on 11 years (1980 to 1991, omitting the year of the forecast) of ECMWF analyses from the observed or predicted value. If $F$, $A$, and $C$ are the forecast, the corresponding analysis, and the corresponding climatology, the mean ACC is calculated by :

$$\mathrm{ACC} = \frac{< \overline{(F-C)(A-C)} >}{\sqrt{< \overline{(F-C)^2} >< \overline{(A-C)^2} >}} \qquad (1)$$

where the overbar indicates the average for the different cases, and the brackets the average over the area. $F$ and $A$ can be instantaneous fields or time averages. Note that with this definition, the mean ACC is not the arithmetic mean of the 32 ACCs (but in our results it is slightly larger). In this paper we shall use only this criterion to evaluate the skill and we shall restrict to the 500 hPa height field.

A positive mean ACC indicates that the forecasts are not independent of the verifying analyses, and thus exhibit some skill. However this skill may be just a potential skill, and, generally, the forecasts need a linear correction to give better results than the climatology forecast ($F = C$) in mean square sense.

Moreover, as pointed out by *Murphy* (1990), the use of a climatology based on a small sample introduces a positive bias in the estimate of the ACC. The impact of this bias on our forecasts has been presented in *Déqué and Royer* (1991). Table 2 recalls the main results. The climatology $C$ has been calculated with $n$ years ($1 \leq n \leq 11$). These $n$ years are chosen at random among the 11 years avalaible for each case (out of the 32), and 100 mean ACCs are calculated from such simulations. The average provides an estimate of the expected mean ACC. One can see on Table 2 that if we use only 5 years to calculate the climatology, the mean ACC is larger than with 10 years by .06 for monthly means. In section 3.5, we shall try to estimate the limit when $n$ tends to infinity, and thus the bias due to the 11-sized sample.

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| day 1-15 | .75 | .69 | .66 | .65 | .64 | .64 | .63 | .63 | .62 | .62 | .62 |
| day 15-29 | .51 | .38 | .31 | .27 | .25 | .24 | .22 | .21 | .21 | .20 | .20 |
| day 30-44 | .42 | .30 | .23 | .20 | .18 | .17 | .16 | .15 | .14 | .14 | .13 |
| day 15-44 | .53 | .39 | .31 | .27 | .24 | .22 | .21 | .20 | .19 | .18 | .17 |

Table 2: Estimates of the mean ACC for 15-day and 30-day means for different sizes $n$ of the sample used to calculate the climatology.

### 3.2 Ensemble mean technique

The forecasts we study are ensemble mean forecasts. They consist of the average of 5 individual forecasts starting at very close situations. This technique allows to reduce the impact of small initial errors which grow during the integration. With this technique, we can also estimate the limit of potential predictability (*Lorenz*, 1982). When the individual forecasts become independent, e.g. when their distance is as large as the distance between forecasts starting at independent situation, one can consider that the model has lost the memory of the initial conditions. This memory of the starting situation is neither a necessary condition of predictability (the boundary conditions can be a source of information for the model), nor a sufficient condition since the model is not perfect and introduces errors in the course of the integration. However, it is an useful tool to estimate a reasonable length for the integrations.
We use here, as a measure of consistency between the individual forecasts, the forecast agreement, introduced by *Murphy* (1990). This coefficient is the average of the 5 ACCs between the ensemble mean forecast and each member of the ensemble. This coefficient is large, because the individual and the mean forecast are not independent. A 95% confidence interval has been estimated for this coefficient. We have generated 5-sized ensembles for which each member is taken at random in different years (but in the same month and at the same forecast lag because of the seasonal cycle and the model drift). In the case of day 30-44 means, this interval is [0.69,0.75]. Figure 1 shows the time evolution of the agreement for 1-, 5-, 10-, and 15-day means and the 95% confidence interval at the time the agreement is no more significant. One can see that instantaneous individual forecasts are significantly dependent up to day 35. Taking time averages increases this lag, and one can see that day 30-44 means are still in agreement.
In a similar way as in section 3.1 we can estimate the impact of the ensemble average on the mean ACC. Here $n$ individual forecasts are drawn out of the 5. Table 3 shows the different estimates. On can see that this impact is very weak, particularly for day 15-44 means. We shall see in section 3.3 that the sensitivity to $n$ is increased when the systematic error is removed.

| $n$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| day 1-15 | .57 | .60 | .61 | .62 | .62 |
| day 15-29 | .17 | .18 | .19 | .19 | .20 |
| day 30-44 | .12 | .13 | .13 | .13 | .13 |
| day 15-44 | .16 | .17 | .17 | .17 | .17 |

Table 3: As Table 2, but for different sizes $n$ of the forecast ensemble.

### 3.3 Systematic error

The climatology of the model being somwhat different from the observed climatology, the model

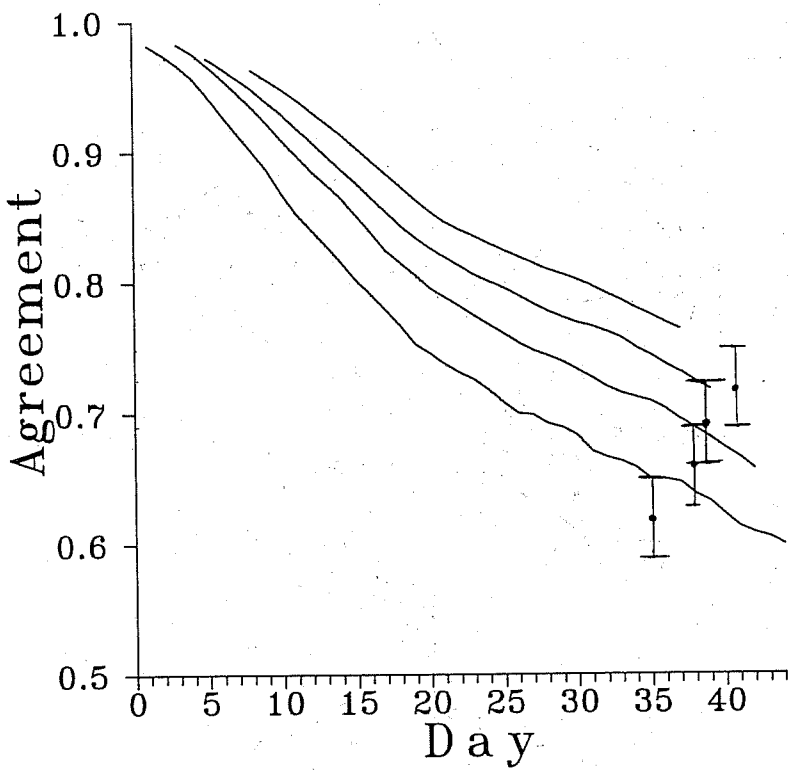


Figure 1: Forecast agreement as a function of the forecast lag for 15-day running means (top), 10-, 5-, and 1-day running means ; 95% confidence interval

forecasts exhibit a systematic drift during the integration. Figure 2 shows the mean error for day 15-44 means. This error consists of 2 negative minima over the Atlantic and the Pacific oceans. It is systematic, since Student t-tests reject the hypothesis of sampling residual to explain this pattern. The model is too cold and this cooling occurs during the first two months of model integration. One can also remark that these two minima correspond to centers of the blocking activity (*Lejenas and Okland*, 1983). In fact this model does not generate blocks over the Atlantic, and generates blocks with half the observed frequency over the Pacific.

There are two ways to reduce this error. One is *a priori* and the other *a posteriori*. We have tried the first way, by using a method similar to *Sausen and Ponater* (1990). The initial temperature drift is estimated at each model level and for the spectral coefficients corresponding to a T10 truncation. This estimate is the time derivative of a 2nd degree polynomial adjusted by least squares from the errors of the first 9 days of a series of reference forecasts. Then the drift is subtracted, at each time step, in the temperature evolution equation (i.e. we introduce an artificial heat source in the model). This method has been tested for 8 January forecasts. Table 4 shows the ACC of the corrected and the uncorrected model. The results are not very encouraging. The mean ACC has been increased, but the skill of the uncorrected model was particularly poor, and the improvement is not systematic (only 3 forecasts are improved). A

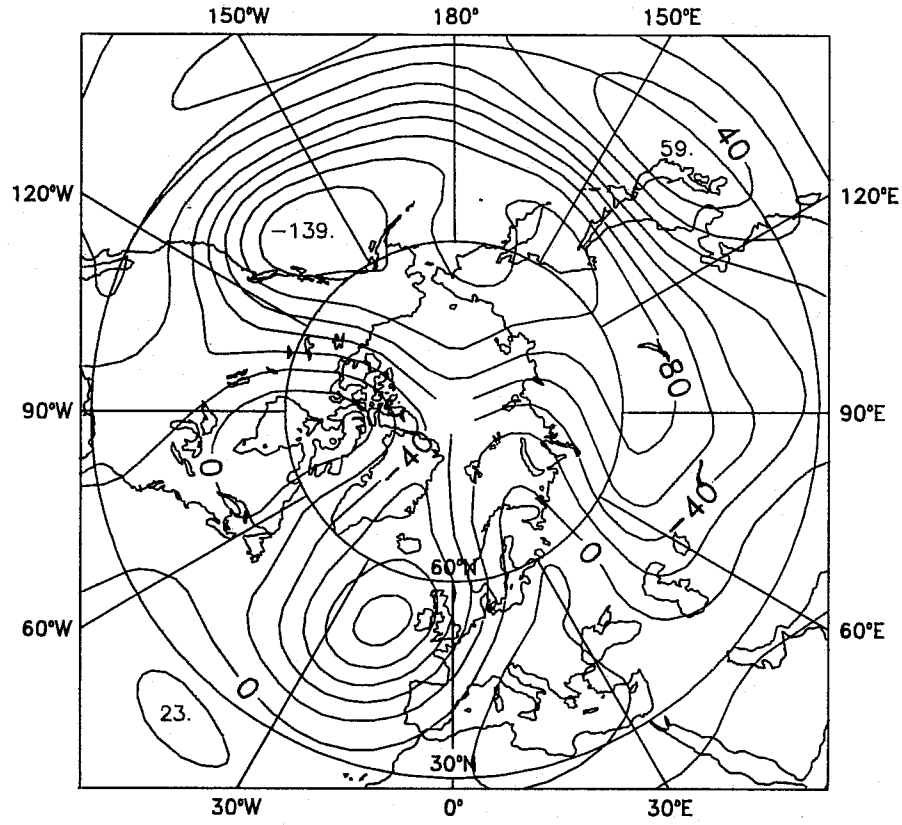


Figure 2: Mean error of the 500 hPa height for day 15-44 means ; contour interval 20 m.

possible improvement of the method could be obtained by calculating the drift in a different way : in the lower troposphere the uncorrected model is warming up during the first 5 days, then cools down ; thus the correction we have used provides an additional cooling at these levels.

| | J.84 | J.85 | J.86 | J.87 | J.88 | J.89 | J.90 | J.91 | Mean |
|---|---|---|---|---|---|---|---|---|---|
| uncorrected | -.14 | -.10 | -.08 | .18 | .35 | .37 | .23 | -.13 | .09 |
| corrected | .16 | .46 | -.09 | .12 | .19 | .48 | .06 | -.21 | .17 |

Table 4: ACC for day 15-44 for the *a priori* corrected and uncorrected forecasts.

The second way to reduce the error consists of subtracting the mean error of the corresponding period (e.g. day 15-44). This estimate must be obtained independently of the current forecast, otherwise the skill is artificially increased. In our case, a mean model error has been calculated for each winter (by averaging the 4 months). For each forecast, the bias which is subtracted to the model results is the average of the 7 biases corresponding to the 7 other winters. Thus this bias does not include any information related to the forecast. With this method the mean

ACC of the day 15-44 means increases from 0.17 to 0.26 and 19 forecasts are improved. The sensitivity to ensemble averaging is better since corrected individual forecasts yield a mean ACC of 0.20.

As in section 3.1 and 3.2, one can estimate the impact of the number of years used to estimate the bias. In *Déqué* (1991), it is shown that when the bias is estimated with a too short sample, the correction decreases the skill. Table 5 shows that the impact of the correction is larger for day 30-44 and 15-44. For shorter ranges the correction is detrimental with $n = 1$.

| $n$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| day 1-15 | .62 | .59 | .62 | .63 | .64 | .64 | .65 | .65 |
| day 15-29 | .20 | .19 | .22 | .23 | .23 | .24 | .24 | .24 |
| day 30-44 | .13 | .15 | .18 | .19 | .20 | .20 | .20 | .21 |
| day 15-44 | .17 | .20 | .22 | .24 | .25 | .25 | .26 | .26 |

Table 5: As Table 2, but for different sizes $n$ of the sample used to calculate the systematic error.

### 3.4 Signification of the scores

Even after correction of the bias, the skill is low. A value of 0.26 for the ACC corresponds to a skill score (as defined by *Murphy and Epstein*, 1989) less than .06. Figure 3 shows the 32 values for the model and persistence ACC. The persistence here is the day −29-0 mean. Different periods have been tried to define the persistence and this one gives the largest ACC (.06 for the mean ACC). The model is clearly superior to the persistence forecast. We count 6 values below 0 and 4 values above 0.50. This large variability explains why the scores reported in the literature may be different. We have performed Monte Carlo simulations of the ACC and the 95% confidence interval for a mean ACC based on 32 independent forecast is [.12,.38]. This interval confirms also the fact that our mean ACC is significantly different from zero, i.e. the model forecasts are not just noise. The same is true for the 15-day means : the intervals for days 1-15, 15-29, and 30-44 are [.60,.69], [.12,.36], and [.11,.29]. One can also perform Monte Carlo simulations with scrambled data : in this case the forecast and the corresponding analysis are related to different years. For day 15-44 means, the 95% confidence interval for the mean ACC when the forecast have no skill at all is [-.02,.20]. This interval is not centered, which confirms the statement of section 3.1. The value of 0.26 lies outside this interval, which confirms that this score is statistically significant.

### 3.5 Extrapolations

In sections 3.1, 3.2, and 3.3 we have estimated by subsampling the mean ACC as a function of

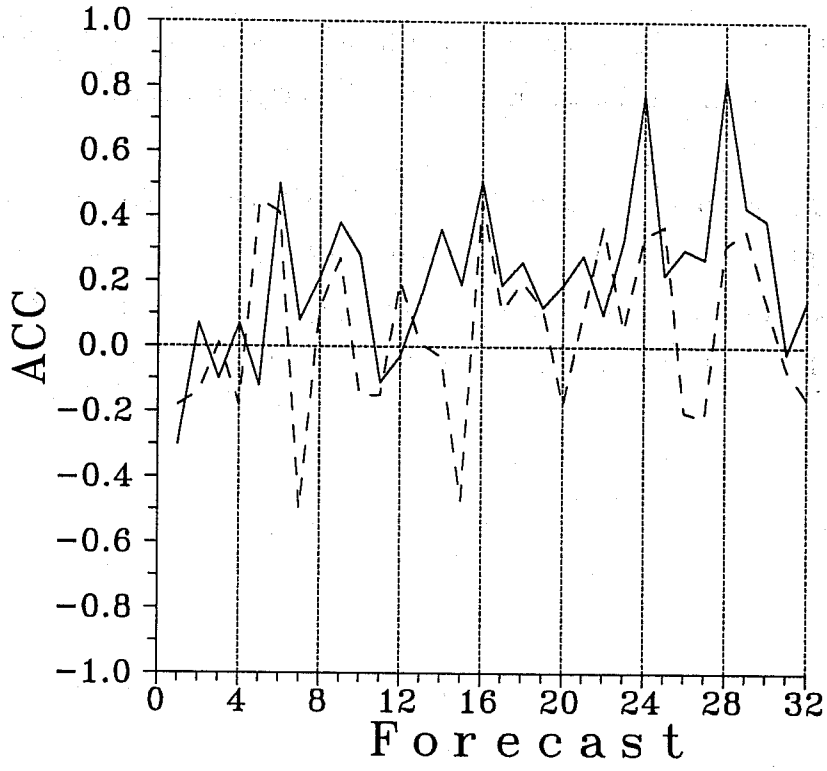


Figure 3: ACC of the model (solid) and the persistence (dashed) for the 32 15-44 forecasts

the sizes of the climatological dataset, of the forecast ensemble, and of the reference dataset. In *Déqué* (1991) and *Déqué and Royer* (1991) we have developed methods to extrapolate the estimates. We present here a simplified method which generalizes the results.

At a given grid point and for a given forecast we have 5 values ($F_1$, ..., $F_5$) the average of which, $F$, is the ensemble mean forecast, the corresponding analysis $A$, 11 values ($C_1$, ..., $C_{11}$) the average of which, $C$, is the climatology, and 7 values ($B_1$, ..., $B_7$) the average of which, $B$, is the bias. Let us consider $n_1$ independent random variables ($\mathcal{F}_1$, ..., $\mathcal{F}_{n_1}$) which have the same statistical distribution as the $F_i$. We introduce similarly the random variables ($\mathcal{C}_1$, ..., $\mathcal{C}_{n_2}$) and ($\mathcal{B}_1$, ..., $\mathcal{B}_{n_3}$). We set the forecast and observed anomaly :

$$X = \frac{1}{5}\sum_{i=1}^{5} F_i \; - \; \frac{1}{11}\sum_{j=1}^{11} C_j \; - \; \frac{1}{7}\sum_{k=1}^{7} B_k \; = \; F - C - B \tag{2}$$

$$Y = A \; - \; \frac{1}{11}\sum_{j=1}^{11} C_j \; = \; A - C \tag{3}$$

Similarly we introduce two random variables :

$$\mathcal{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathcal{F}_i \; - \; \frac{1}{n_2}\sum_{j=1}^{n_2} \mathcal{C}_j \; - \; \frac{1}{n_3}\sum_{k=1}^{n_3} \mathcal{B}_k \tag{4}$$

$$\mathcal{Y} = A - \frac{1}{n_2}\sum_{j=1}^{n_2} C_j \tag{5}$$

A random mean ACC obtained with an ensemble size $n_1$, a climatology based on $n_2$ years, and a bias estimated with $n_3$ years is :

$$\mathcal{ACC} = \frac{<\overline{\mathcal{XY}}>}{\sqrt{<\overline{\mathcal{X}^2}><\overline{\mathcal{Y}^2}>}} \tag{6}$$

To estimate the statistical expectation of $\mathcal{ACC}$, we set :

$$\mathcal{X}^2 = E(\mathcal{X}^2) + \epsilon_1 \quad \mathcal{Y}^2 = E(\mathcal{Y}^2) + \epsilon_2 \quad \mathcal{XY} = E(\mathcal{XY}) + \epsilon_3 \tag{7}$$

E() is the expectation and $\epsilon_1$, $\epsilon_2$, and $\epsilon_3$ are centered random variable. One can neglect $<\overline{\epsilon_1}>$, $<\overline{\epsilon_2}>$, and $<\overline{\epsilon_3}>$. This approximation is better when $n_1$, $n_2$, and $n_3$ are large. However the approximation is not exact, even when they tend to infinity because we use for the expectations unbiased but rough estimates :

$$\begin{aligned} E(\mathcal{X}^2) \simeq\ & X^2 + (\frac{1}{n_1} - \frac{1}{5})\{\frac{1}{4}\sum_{i=1}^{5} F_i^2 - \frac{5}{4}F^2\} \\ & + (\frac{1}{n_2} - \frac{1}{11})\{\frac{1}{10}\sum_{j=1}^{11} C_j^2 - \frac{11}{10}C^2\} \\ & + (\frac{1}{n_3} - \frac{1}{7})\{\frac{1}{6}\sum_{k=1}^{7} B_k^2 - \frac{7}{6}B^2\} \end{aligned} \tag{8}$$

$$E(\mathcal{Y}^2) \simeq Y^2 + (\frac{1}{n_2} - \frac{1}{11})\{\frac{1}{10}\sum_{j=1}^{11} C_j^2 - \frac{11}{10}C^2\} \tag{9}$$

$$E(\mathcal{XY}) \simeq XY + (\frac{1}{n_2} - \frac{1}{11})\{\frac{1}{10}\sum_{j=1}^{11} C_j^2 - \frac{11}{10}C^2\} \tag{10}$$

In the case of day 15-44 means, we get the formula :

$$E(\mathcal{ACC}) \simeq \frac{.18 + 2.04/n_2}{\sqrt{(2.10 + 2.04/n_2)(.48 + .76/n_1 + 2.04/n_2 + .62/n_3)}} \tag{11}$$

This formula is exact by construction for $n_1 = 5$, $n_2 = 11$, and $n_3 = 7$. A good test of accuracy consists of letting $n_1 = 1$, $n_2 = 1$, or $n_3 = 1$ and to compare with the first columns of Table 2, Table 3, and Table 5. The differences are less than .01, so we can be rather confident in the approximation, and try to use the formula for extrapolation. When the size of the ensemble tends to infinity, the mean ACC tends to 0.28 (instead of 0.26 when $n_1 = 5$). The impact of a larger ensemble on the mean skill is thus negligible. When the size of the climatological

dataset tends to infinity, the mean ACC falls down to 0.15. The bias is thus as large as 0.11; this result agrees with that of section 3.4 : after scrambling the forecasts and the analyses, the mean ACC is, on the average, 0.09. The impact of the size of the correction dataset is of the same order as that of the ensemble size (but each year of this dataset corresponds to 20 individual integrations of the model) : when it tends to infinity the mean ACC tends to 0.27.

## 4. SKILL VERIFICATION

### 4.1 Standard predictors

As shown in Fig.3, the score exhibits large case-to-case fluctuations. If we should be able to discriminate *a priori* the good and the bad forecasts, the mean model skill would be improved by using the model output only in the cases when the model is expected to be successful. Another possible use could be to better document the forecast by providing a quality index together with the forecast. The ultimate use of skill forecast is the probability forecast in which our confidence in the model results is expressed in term of probability distribution. In this section 4 we shall restrict to day 15-44 means.

A classical candidate for predicting the skill is the model spread, i.e. the standard deviation of the forecast ensemble (*Kalnay and Dalcher*, 1987). The basic idea is that when the trajectories diverge, there is less predictability. In fact, it is better to take a normalized index: if the 5 individual forecasts predict a strong positive anomaly over a large area, but with different intensities, the spread may be large, but we are relatively confident in this forecast. We have taken the ensemble agreement as in section 3.2. Table 6 shows that this index is better correlated than the spread with the ACC. However the correlations are not statistically significant : we have estimated the 2.5 and the 97.5 percentile of the distribution of correlation when there is no forecast skill (and consequently no predictability of skill), in a similar way as in section 3.4. Among the 14833 permutations of the 8 years which have no coincidence (i.e. $n(i) \neq i$, $i = 1,\ldots,8$), 1000 permutations are taken at random, and the forecast dataset is scrambled according to each permutation. We have thus 1000 series of 32 ACCs and 32 spreads (or any other predictor of the ACC). The percentiles are computed from the 1000 correlation coefficients. The stability of the method is verified by repeating the process with 1000 other permutations.

Two indices however seem significantly correlated with the ACC. The first one is the amplitude of the forecast anomaly (measured by its spatial root mean square). The fact that the mean ACC calculated by Eq. (1) is larger than the arithmetic average of the ACCs suggests that

| index | SPR | AGR | AMP | PER |
|---|---|---|---|---|
| $CC$ | .09 | .21 | .31 | .25 |
| $CC_{2.5}$ | -.25 | -.45 | -.45 | -.39 |
| $CC_{97.5}$ | .25 | .28 | .28 | .21 |

Table 6: Correlation coefficient ($CC$) between the ACC and various indices : root mean square error (RMS), ensemble spread (SPR), ensemble agreement (AGR), forecast anomaly amplitude (AMP), anomaly correlation between model and persistence (PER). The values $CC_{2.5}$ and $CC_{97.5}$ indicate that 95% of the $CC$ belong to the interval $[CC_{2.5}, CC_{97.5}]$ when the forecast are scrambled.

the cases with large forecast or observed amplitude have a larger ACC than the others. The positive correlation between amplitude and ACC does not result of algebraic relations since, after scrambling, the correlation tends to be negative. We can also remark that the amplitude takes into account the agreement between the individual trajectories : when the 5 integrations are very different, their contributions to the ensemble mean compensate each other and the amplitude is weak. This is verified by the fact that the correlation between amplitude and ACC for the individual forecasts is only 0.21.

A second predictive index is the anomaly correlation between forecast and persistence (i.e. day $-29$-0 mean). This index is calculated in the same way as the ACC, just replacing the observed by the persistent anomaly. The choice of this index comes from a feature of Fig.3 : the good model forecasts are also good persistence forecasts. High values of this index correspond to situations for which the large-scale patterns have not been modified during the integration. The anomaly correlation is preferable to RMS differences, since the impact of amplitude counteracts the impact of persistence. Using 1-day, 5-day or 10-day means instead of 30-day means for the persistence leads to positive, but smaller correlation coefficients.

One can also remark on Table 6 that the confidence intervals are not centered. This feature is stable if we take other sets of permutations. The following two tests illustrate the fact that this is a consequence of the way of calculating the anomaly. The same method is applied for independent gaussian variables (assuming the area is a single point, the forecast and observation are independent, and the variance of the observation is 4 times that of the forecast). If we simulate the subtraction of the estimated climatology and of the estimated bias, the 95% interval for the correlation between AMP and ACC is [-.36,.25]. If we directly generate independent forecast and observed anomalies, the interval is [-.36,.36], and the expectation of the correlation can be analytically calculated (it is simply 0).

4.2 Composites according with the skill

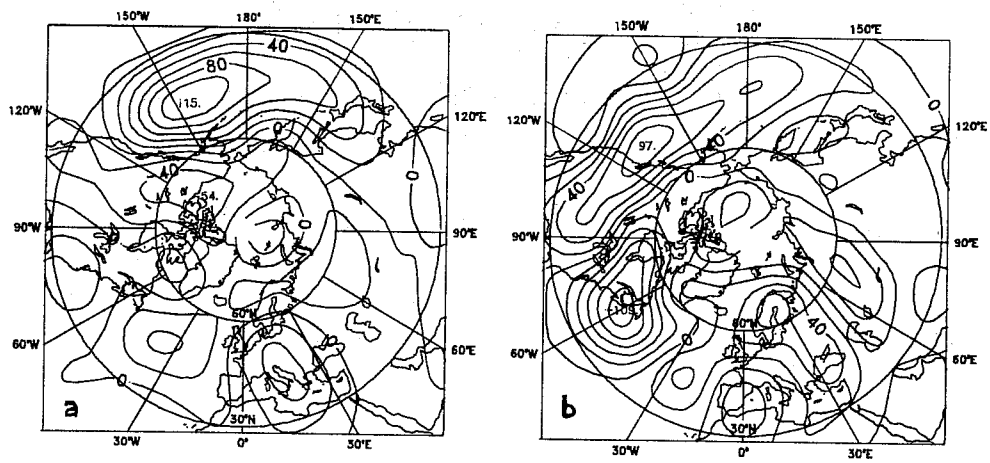


Figure 4: Difference between the composite maps of the 16 best and the 16 worst forecasts ; day 15-44 (a) and day −4-0 (b) mean analyzed anomalies ; contour interval 20 m.

Figure 3 shows that, among the 32 forecasts, 16 have an ACC above .20 and 16 below .20. We have calculated composite maps of several fields by averaging the 16 best cases or the 16 worst cases. Figure 4a shows the difference between the best and the worst cases for the day 15-44 analyzed anomalies. A similar map (not shown) of the t-values shows that the positive area over the Pacific is significant at the 95% level. The good forecasts correspond to positive anomalies over this area. However this method is an *a posteriori* one since the analysis is unknown at the time of the forecast. If we perform the same computation for the forecast anomalies, the best and worst forecasts do not exhibit differences at the 95% level.

One can try to investigate the impact of initial conditions on the skill. In the medium range *Palmer* (1988) found that a negative PNA was associated with a bad skill. He explained it by the fact that the atmospheric situation was more unstable in this regime. We have produced composite maps as Fig. 4a, but for situations before the starting date. We have tried day 0, day −4-0, day −9-0, and day −29-0 mean anomalies. The patterns are similar, and the highest significance in t-tests is obtained for 5-day means. Figure 4b displays this pattern : only the dipole eastern coast-western coast of America is significant. This pattern is different from the PNA. However, if the same analysis is performed, using the mean square error of day 1-15 mean forecasts for the discrimination, the map corresponding to Fig. 4b (not shown) exhibits a positive PNA.

### 4.3 Probability forecasts

Since a large uncertainty is attached to the model forecasts, it is natural to try to express them as probabilities. In the case of the 500 hPa height field, the best way is to attempt to give probabilities to different typical clusters as in *Brankovic et al.* (1990). However, for

the sake of simplicity, we adopt here a local formulation, which should be rather adapted to fields like temperature (see *Déqué*, 1988b). At each grid point the forecast is a gaussian variable ; the mean ($M$) is the ensemble mean forecast and the standard deviation ($\sigma$) is a value which expresses our uncertainty. A natural candidate for the standard deviation is the ensemble spread. Then, a set of $n$ probabilities $p_i$ is given for $n$ categories. These categories are equiprobable with respect to climatology, i.e. the thresholds are calculated from a gaussian distribution estimated with the 11 years available (which depends on the calendar month). The forecast will be excellent when the $p_i$ will be close to 0, except $p_j \simeq 1$, where the $j$th category is observed. Let $o_j = 1$ when the $j$th category is observed and $o_j = 0$ otherwise. A criterion which measures the skill of such a forecast is the ranked probability score (RPS) introduced by *Epstein* (1969) :

$$\mathrm{RPS} = \sum_{i=1}^{n} \left( \sum_{k=1}^{i} (p_k - o_k) \right)^2 \tag{12}$$

This RPS can be averaged for the northern hemisphere and for the 32 forecasts. Similarly one can introduce the RPS of the climatological forecast $\mathrm{RPS}_c$. This forecast consists of taking $p_i = 1/n$. It minimizes the mean RPS in the absence of any skill (i.e. when $o_i$ and $p_i$ are independent). A skill score can be introduced as :

$$SS = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS}_c} \tag{13}$$

Table 7 displays the values of $SS$ for 2, 3, 5, and 101 categories. The case of 101 categories is a close approximation of the continuous case (see *Déqué*, 1988b). The first row correspond to the deterministic forecast ; here $\mathrm{RPS}_c$ is calculated differently : the deterministic climatology forecast consists of taking 1 for the medium category ($n$ must be odd) and 0 for the others. The "natural" choice of taking the ensemble spread for the standard deviation of the probability distribution leads to a very bad score (2nd row). Two reasons explain this : as seen in section 4.1, the spread is not correlated with the skill ; moreover the spread measures the uncertainty due to the initial error increase, and thus underestimates the total uncertainty since the model is not perfect. Simply averaging the 32 spreads (3rd row), assuming no relation between the fluctuations of the skill and the fluctuations of the spread, reduces the RPS. Different combinations of spread or amplitude have been tried to choose a standard deviation which maximizes the $SS$. Finally the different attempts have lead to the same values as those of the 4th row : taking the climatological standard deviation (according to each calendar month) is the best we can do. This imply that we are not able at this stage to predict the standard

deviation (though we can predict the mean). This negative value for the skill score, which do not depend on $n$, is of the same order as the skill score of deterministic forecasts. It can be compared with a skill score of $-.07$ obtained as by Eq. (13) but with the mean square error (MSE) instead of the RPS.

| $n$ | 2 | 3 | 5 | 101 |
|---|---|---|---|---|
| $\sigma = 0$ | und. | -.05 | -.01 | -.05 |
| $\sigma$ = spread | -.20 | -.20 | -.18 | -.18 |
| $\sigma$ = mean spread | -.18 | -.17 | -.16 | -.16 |
| $\sigma$ = climatology | -.05 | -.06 | -.06 | -.06 |
| conditional probability | .05 | .05 | .05 | .05 |

Table 7: Skill scores of $n$-category probability forecast for different choices of the probability distribution.

We can try to improve the forecast if we do not impose to take the ensemble mean forecast for $M$. When there is no skill at all, the probability distribution which minimizes the expectation of the RPS is the climatological one. When forecast and observation are not independent, it can be demonstrated that the expectation of the RPS is minimized by taking the conditional distribution of the observed anomaly $\alpha_o$ for a given forecast anomaly $\alpha_f$. Let us assume that, at a given grid point, the pair ($\alpha_o$,$\alpha_f$) is a gaussian vector of means (0,0), standard deviations ($\sigma_o$,$\sigma_f$), and correlation $r$. When $\alpha_f$ is set to $F - C - B$, the variable $\alpha_o$ is a gaussian of mean $r\sigma_f/\sigma_o(F - C - B)$ and standard deviation $\sqrt{1 - r^2}\sigma_o$. When $r$ is 0 (no skill) , this distribution is the climatological one. When $r$ is 1 (excellent forecast), we get the deterministic forecast. Here $r$ is of the order of .3 and the ratio of the standard deviations of 2. Thus the forecast anomaly is reduced by a factor of about .6, and the climatological standard deviation is practically unchanged. Table 7 shows that with this correction the skill score becomes positive. Note that, with such a correction, the skill score of deterministic forecasts based on the MSE is .06 (see section 3.4).

## 5. CONCLUSION

We have studied the skill of time averaged forecasts through the northern hemisphere 500 hPa height anomaly correlation. This skill is significantly different from 0 and is better than the skill of persistence forecasts. However this skill is very low and the model forecasts need to be damped to outperform the climatology forecasts. The scores are improved by ensemble average and bias correction : a minimum size for the ensemble is 3 individual integrations, and for the

reference dataset used to estimate the bias is 3 winters ; with larger sizes the score increase is negligible. The scores are overestimated by the use of a climatology based on 11 years ; however the use of larger sample (e.g. 30 years) could introduce inhomogeneities and finally overestimate the scores ; when the *Oort* (1983) climatology based on 1963-1973 observations is used as reference, the scores are larger by about .15.

The skill exhibits a strong case to case variability, and even 8-winter mean scores can fluctuate by plus or minus .10. Two possible predictors of this variability are found, though their statistical significance is marginal. The first one is linked to the size of the anomaly, the second one to the difference between the initial and the final state of the forecasts : when the model predicts a strong and persistent anomaly, the skill is larger. There are also significant connexions between the skill and the initial situation, but they do not correspond to those obtained in the short and medium range. The model forecast can be improved by a local probability formulation, but the best trivial forecast (i.e. the climatology) is improved too and the relative skill remains unchanged.

## References

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

Bougeault, P., 1985: A simple parameterization of the large-scale effects of deep cumulus convection. *Mon. Wea. Rev.*, **113**, 2108–2121.

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

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

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

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

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

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

Déqué, M. and J.F. Royer, 1991: The choice of an observed climatology to verify extended range forecasts. In *Extended abstracts submitted to the ICTP/WMO international technical conference on long-range weather forecasting research, Trieste, 8-10 April 1991*, pages 73–76, WMO/TD No 395.

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

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

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

Lejenas, H. and H. Okland, 1983: Characteristics of northern hemisphere blocking as determined from a long time series of observational data. *Tellus*, **35A**, 350–362.

Lorenz, E.N., 1982: Atmospheric predictability with a large numerical model. *Tellus*, **34**, 505–513.

Murphy, A.H. and E.S. Epstein, 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.

Oort, A.H., 1983: *Global atmospheric circulation statistics, 1958-1973.* Technical Report, NOAA professional paper no 14, US Government Printing Office, Washington, DC 20402, 180 pp.

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

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

Sausen, R. and M. Ponater, 1990: Reducing the initial drift of a gcm. *Beit. Phys. Atmos.*, **63**, 15–24.

Simmons, A.J. and D.M. Burridge, 1981: An energy and angular momentum conserving vertical finite-difference scheme and hybrid vertical coordinate. *Mon. Wea. Rev.*, **109**, 758–766.

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

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