Tellus (1997), 49A, 74–86
Printed in UK – all rights reserved

Copyright © Munksgaard, 1997
TELLUS
ISSN 0280–6495

Ensemble size for numerical seasonal forecasts

By M. DÉQUÉ, *Météo-France, Centre National de Recherches Météorologiques 42 Av. Coriolis, Toulouse, France*

(Manuscript received 8 January 1996; in final form 17 May 1996)

ABSTRACT

The predictability of 500 hPa height, 850 hPa temperature and precipitation is studied using the "perfect model" approach with ensemble numerical forecasts. The sea surface temperature interannual variability is introduced to provide some source of seasonal predictability. The mean scores of 16 winter forecasts show a large potential in the tropics, a weak one in the midlatitudes, in particular over Europe. The weakness of the scores in the midlatitudes may be partly explained by the underestimation of the amplitude of the seasonal anomalies by the model. The seasonal forecasts are based on 9 individual model integrations starting at slightly different initial conditions. The variation of the scores with the size of the ensemble is estimated empirically. It is shown that, for a perfect seasonal forecast, the size necessary to approach the saturation score is about 3 for the tropical precipitation, 20 for midlatitude height, and 40 for temperature over Europe.

1. Introduction

The use of comprehensive models of the atmosphere has led to important progress in short- and medium-range forecasting. The integration of such models one month ahead has demonstrated some skill (Royer, 1993), but, due to the weak part of the informative signal with respect to the noise, the forecast is no longer deterministic, and a probability formulation has to be found (Déqué et al., 1994a). A way to filter out the unpredictable noise, or, equivalently, to take into account the different weather regimes which may occur, is to consider long time averages. In recent years, there has been increasing interest in seasonal-scale predictions (Palmer and Anderson, 1994; Stern and Miyakoda, 1995). Extending model integrations to such ranges is now possible by the considerable progress in ocean-atmosphere coupled models.

Averaging meteorological fields for a whole season is not sufficient to remove the random component, as shown by Brankovic and Palmer (1996), in particular in the midlatitudes. The ensemble average technique, which has proved its efficiency in monthly forecasting (e.g. Brankovic et al., 1990) is also necessary for seasonal forecasting.

Recently, a coordinated experiment has started, involving the United Kingdom Meteorological Service (UKMO), the European Centre for Medium-range Weather Forecast (ECMWF), the French electricity company (EDF) and the French meteorological service (Météo-France). Its aim is to evaluate with some accuracy the potential for seasonal forecasting. This experiment consists of integrating several models for all the seasons of the 1979–1994 period, using the US National Meteorological Center (NMC) analyses of sea surface temperature (SST) as boundary conditions (the ocean is thus assumed to be perfectly predicted), and the land surface and atmosphere data from the ECMWF re-analysis (ERA) as initial conditions. Thus, 64 120-day forecasts are to be produced by each of the 4 participants (only 16 in the case of EDF which is mainly interested in winter forecast). Each forecast consists of an

Corresponding author. e-mail: deque@meteo.fr

ensemble of 9 24-h lagged model integrations; the lagged average technique (Hoffman and Kalnay, 1983) is a convenient way to produce Monte Carlo forecasts, since the starting situations for each case are equilibrated with respect to the model and easy to obtain. This method does not take into account the uncertainty about the initial state because the best agreement between the lagged forecasts occurs in data void areas where the analyses are model forecasts. However, since the initial errors grow rapidly (Stroe and Royer, 1993; Savijärvi, 1995) and saturates after about 20 days in the troposphere, the technique used to perturb the atmospheric initial state has little importance. This experiment is expected to finish after the completion of the ERA project (end of 1996) and will be referred to as IAPO (imperfect atmosphere–perfect ocean). Further experiments are planned with a fully coupled ocean-atmosphere system.

The objective of this study is to evaluate an estimate of a reasonable ensemble size. The choice of 9 members in IAPO is dictated by the computational cost of such an experiment (it is equivalent to about 200 years of simulation for each participant). One can consider that the error in a seasonal forecast comes from two distinct sources: the fact that the model does not follow the exact laws of the atmosphere, and the small unavoidable initial errors which increase due to the model nonlinearities. Ensemble averaging has no impact on the first source, but reduces the random component in the model results. It thus enhances the model response to the variables which are prescribed (SST for an uncoupled model) or which evolve slowly (soil moisture, snow cover, QBO phase, etc). Even if the model is perfect, the mean forecast does not converge toward the verification as the ensemble size tends to infinity, but to a statistical situation with smaller anomalies than the observed data: the actual atmosphere "chooses" one trajectory, and the ensemble mean forecast gives the expectation of all possible trajectories. With a given ensemble size, the forecast error is the sum of the model error and the inaccuracy of the ensemble mean:

$$\bar{F}-V=(E(F)-V)+(\bar{F}-E(F)), \qquad (1)$$

where V is the verification, $\bar{F}$ the ensemble mean and $E(F)$ the limit of $\bar{F}$ as the ensemble size tends to infinity. When the model is perfect, the first term of eq. (1) is not zero since V comes from a single trajectory, and $E(F)$ is an average of a large number of possible values for V; but it is generally smaller than when the model is not perfect, since a perfect model has no systematic error. The second term is related to the model spread, since it can be written as the ensemble average of the deviations $F-E(F)$; it decreases toward zero as the ensemble size tends to infinity.

In this study the impact of the ensemble size on the score of a "perfect model" is investigated. A good choice for the ensemble size makes the second term of eq. (1) smaller than the first one. Since the first term is generally smaller with a perfect model than with an imperfect one, a reasonable size in "perfect model" will be a fortiori good for actual forecasts. The perfect model approach is very popular in predictability studies (Barker, 1991 or Houtekamer and Derome, 1994). It consists of using one particular integration of an ensemble of model simulations starting from similar situations as the verification trajectory, the other integrations forming the forecast ensemble. In the present study, we introduce an ingredient to counteract the loss of predictability due to the divergence of the trajectories: the integrations use year-to-year varying SSTs, the same for each member of a given ensemble.

The perfect model experiment, referred to as PAPO (perfect atmosphere-perfect ocean) is described in Section 2. In Section 3, the scores of the "perfect model" are given for a few parameters and time ranges; estimates of the scores as a function of the ensemble size are derived from Monte Carlo drawings and from statistical calculations in Section 4. Concluding remarks, including perspectives about the "imperfect model", are made in Section 5.

2. Description of the model and the experiment

The atmospheric GCM used for this study is a climate version of ARPEGE-IFS (cycle 12). This model has been developed by Météo-France and ECMWF for operational short- and medium-range forecast. The climate version derived from this code is described in Déqué et al. (1994b). The version used in the PAPO experiment is the same as the one used in IAPO. The model used by EDF in IAPO is the same version, except the horizontal resolution (T63). The models used by ECMWF

in IAPO and in ERA are also versions of ARPEGE-IFS, but with a different set of physical parameterizations.

The model has a T42 spectral truncation. The horizontal resolution used for the non-linear terms of the Navier-Stokes equations and for the physical parameterizations is about 300 km. The model has 31 vertical levels from the earth surface to about 80 km. The 20 upper levels are constant in pressure and located above 200 hPa. This allows a good representation of the stratosphere dynamics. The physical parameterizations in the stratosphere are the radiation (from Geleyn and Hollingsworth (1979)), the orographic gravity wave deposition, and a linearized photochemical scheme (from Cariolle and Déqué (1986)) to calculate the ozone sources and sinks. The ozone concentration is a three-dimensional prognostic variable coupled to the rest of the model by the radiation and the advection. In the troposphere, the physical parameterizations also include the convection (a mass flux scheme with Kuo closure from Bougeault (1985)) and the large-scale precipitation. The cloudiness is diagnosed at each tropospheric level from the relative humidity and the convective activity through empirical formulae. In the boundary layer (surface and lowest 4 levels) the vertical diffusion is calculated from the Louis (1979) scheme; this scheme includes a shallow convection parameterization and is extended up to the stratosphere with a decreasing mixing length. At the surface, the ISBA (Noilhan and Planton, 1989) soil-vegetation scheme is used to provide a boundary condition to temperature and moisture. Three layers (surface, deep, and climatological) are used for temperature and four reservoirs (canopy, snow, surface, and root) are used for moisture. The time step is 15 min for the dynamics and the physics (including the radiation). The horizontal diffusion uses a ∇^6 spectral damping, with an e-folding time of 9 h for the highest wave number (i.e., $n=42$; this time decreases linearly with pressure above 100 hPa and is divided by 3 for the divergence field).

This model has been integrated from September 1978 through May 1995 using monthly observed SSTs from NMC analyses. The end of 1978 is not included in the study, since the model needs a few months to drift toward its own climate (climatological SSTs have been used for this period). The interannual SST variability is thus taken account in the same way as in ERA. Once this reference simulation has been produced, "pseudo-forecasts" have been carried out in the same way as IAPO, except that only winter forecasts have been performed. The 16 ensembles start from 1 December 1979, 1980, ..., 1994 of the reference simulation. Each ensemble consists of 9 120-day simulations with the same model and the same SSTs. The use of lagged integrations is not possible in a perfect model framework, since the forecasts would be the exact reproduction of the reference. The following procedure has been used to simulate the initial errors. Nine sets of perturbations for each state variable (wind, temperature, moisture, ozone and logarithm of surface pressure) by linear combinations of the 16 "1 December" situations have been produced. The weights w_{ij}, $i=1, 9$, $j=1, 16$ are drawn at random, and normalized so that:

$$\sum_{j=1}^{16} w_{ij}=0, \qquad \sum_{j=1}^{16} w_{ij}^2=0.01\,. \tag{2}$$

The resulting 500 hPa height initial root mean square error (RMSE) over the northern hemisphere is about 10 m, which is much less than the mean initial error in the IAPO experiment (80m), but in better agreement with measurement uncertainties (the 24 h forecast error is about 20 m with the latest forecast of IAPO). The perturbed initial states are to first order in equilibrium with the model.

We have thus 16 independent cases of winter forecasts consisting of one "verification" and 9 individual "forecasts". With the "verification" being obtained by the same process as the 9 "forecasts", it will be considered in the following that 10 members are available, one member taken at random being the "verification", the other 9 forming the "forecast ensemble". Repeating this procedure will increase the statistical stability of the scores given below. As mentioned in Section 1, this experiment will be referred to as PAPO.

3. The scores

Because of the very large amount of data produced by the PAPO experiment, we restrict the analysis to 3 fields and 3 periods. The fields retained are:

- the 500 hPa height over the northern hemisphere (20°N–80°N); this field is the most widely used in predictability studies, since it

synthetizes the general circulation in the midlatitudes;

- the precipitation over the tropical belt (30°S–30°N); this field is rarely used because of the lack of verification data, but is interesting since the main source of predictability in PAPO comes from the SST variability in this region; moreover, the tropical precipitation has a strong impact on human activities;
- the 850 hPa temperature over Europe (12.5°W–42.5°E, 35°N–75°N); the predictability of such a field in winter has some interest for potential users; the advantage of 850 hPa versus surface or screen level is that such a level is independent of the model orography and allows a better intercomparison or hybridation between different models.

The 3 periods retained are:

- the week (day 7–14)
- the month (day 15–44)
- the season (day 31–120)

Here we are mostly interested in the seasonal forecasts, but it is instructive to compare the impact of the length and the lag of different averaging periods. A question one may ask is whether the larger length (90 days) of the season versus the month compensates the larger lag (75 days) between the mid-period and the starting date. It is also of interest to see to which extent the ensemble size can be reduced when using a longer averaging period.

The useful information to extract from the model fields is the quality of the forecasts. Different scores may be used for the verification. The simplest to apply is the RMSE, but it must be scaled by the corresponding value of the climatological forecast, yielding the "Skill Score" (Murphy and Epstein, 1989). This score is pessimistic in seasonal forecasting, since the climatology forecast is generally better in least square sense, although the model forecast contains some information; with an appropriate scaling of the anomalies the model forecast may be made better than the climatology (Déqué and Royer, 1992), but this can appear to be artificial. Another score is the temporal correlation between a series of forecasts and corresponding verifications. But this score cannot be used for a single forecast-verification pair, and the estimates with short series are not very stable from a statistical point of view. The most popular score in meteorology is the Anomaly Correlation Coefficient (ACC), which reads:

$$\mathrm{ACC} = \frac{\langle (F-\mathscr{C})(V-\mathscr{C}) \rangle}{\sqrt{\langle (F-\mathscr{C})^2 \rangle \langle (V-\mathscr{C})^2 \rangle}}, \tag{3}$$

where F is the forecast, V the verification, $\mathscr{C}$ a climatology and $\langle\ \rangle$ the spatial average operator. The problem with this score comes from the choice of the climatology $\mathscr{C}$: the ACC is overestimated when $\mathscr{C}$ is too smooth, and also when it is too noisy. If the climatology is "far" from both verification and forecast, the ACC is artificially high, since the ACC can be interpreted geometrically as the cosine of the angle forecast-climatology-verification in the phase space. Such artificially high ACCs can occur, for example, if one verifies precipitation forecasts against satellite-derived data using a rain-gauge based climatology, or if one verifies surface temperature forecasts using a smooth climatology which does not include the mountains resolved by both the model and the verification. On the other hand, if the climatology is calculated from a short sample, $\mathscr{C}$ includes noisy residuals which are present in both $(F-\mathscr{C})$ and $(V-\mathscr{C})$ and which increase artificially the ACC. A simple way to calculate $\mathscr{C}$ is to take the time average of the verifications, excluding the verification itself for each year. The time average must be calculated with more than 10 years (Déqué, 1991) to limit the overestimation. Even with 16 years, the bias remains perceptible: the ACC of 500 hPa height seasonal means has been calculated by scrambling the forecast series to simulate a forecast without any skill. When repeating the process 100 times, the mean ACC is 0.11.

A way to avoid this bias is to consider two "climatologies", one for the forecast, and one for the verification, as in the classical correlation coefficient. If we modify eq. (3) by:

$$\mathrm{ACC}' = \frac{\langle (F-[F])(V-[V]) \rangle}{\sqrt{\langle (F-[F])^2 \rangle\ \langle (V-[V]^2 \rangle}}, \tag{4}$$

where $[F]$ is the average of the remaining 15 forecasts and $[V]$ the average of the remaining 15 verifications, we get an unbiased score: the mean ACC′ after scrambling as above is -0.01. Eq. (4) is unchanged if the time averages involve the 16 years since:

$$F_i - \frac{1}{m-1} \sum_{j \neq i} F_j = \frac{m}{m-1} \left(F_i - \frac{1}{m} \sum_{j=1}^{m} F_j \right). \tag{5}$$

If the model is not perfect, the ACC' is equivalent to the ACC of the forecast after removing the model bias (Déqué, 1991). This score can be used with only two forecasts as in Brankovic and Palmer (1996) where the model anomaly 1987–1988 is compared to the corresponding observed anomaly.

The aggregation of the ACC' for different years can be done by averaging over all years considered. In this study the 3 covariances that appear in eq. (4) are averaged separately:

$$C_{FV}=\langle(F-[F])(V-[V])\rangle,$$

$$C_{FF}=\langle(F-[F])^2\rangle, \qquad C_{VV}=\langle(V-[V])^2\rangle. \quad (6)$$

Such an aggregated score has been used in Déqué and Royer (1992) with the name of Mean Anomaly Correlation (MAC). It has the advantage of a better interpretation of potential predictability: 1-MAC^2 is the ratio of the model mean square error by the climatology mean square error, after correction of the systematic and amplitude errors. It can be also seen as an average of the different ACC' with a larger weight when a strong anomaly is predicted and observed. Other advantages (as seen below) include the possibility of unbiased statistical averages, and allowing a unified approach with the temporal correlation. This method of averaging covariances, for which unbiased estimates exist, rather than averaging correlations, is natural in statistics: statisticians prefer to average variances than standard deviations, since the latter are estimated with some bias.

The unified scoring approach used in this study is based on a single coefficient, depending on time i ($i=1, 16$) and space x:

$$C(x,i)=\frac{\langle E\{(F(x,i)-[F(x)])(V(x,i)-[V(x)])\}\rangle}{\sqrt{(\langle E\{(F(x,i)-[F(x)])^2\}\rangle\ \langle E\{(V(x,i)-[V(x)])^2\}\rangle}}, \quad (7)$$

where $E\{\ \}$ represents the statistical expectation when, for example, the verification is drawn at random among the 10 members, or when the forecast is drawn at random by scrambling. This coefficient will be referred to as C-score. The aggregation rule is that the 3 covariances are averaged separately, and the C-score is reconstructed from the 3 averages. The 3 averages enable calculation of the RMSE:

$$\mathrm{RMSE}=\sqrt{C_{FF}+C_{VV}-2C_{FV}}. \quad (8)$$

The RMSE of the climatological forecast being simply $\sqrt{C_{VV}}$. The Skill-Score introduced by Murphy and Epstein (1989) is then:

$$\mathrm{SS}=\rho(2C-\rho) \qquad \text{with} \qquad \rho^2=C_{FF}/C_{VV} \quad (9)$$

When rescaling the forecast anomaly, C is unchanged, but ρ may be chosen arbitrarily; the maximum value for SS is thus C^2.

When the C-score is aggregated with respect to i, one obtains the classical temporal correlation. When it is aggregated with respect to x (i.e., over a geographical domain), one obtains the above ACC'. Finally, one can synthetize the skill by aggregating with respect to both space and time: the C-score is then a time-aggregated anomaly correlation, or equivalently a space-aggregated temporal correlation. This C-score has zero mean when there is no skill at all (as shown above with the scrambling procedure). There is no artificial systematic skill, but the confidence interval about zero may be large when the size of the time average is small. Table 1 shows the confidence intervals for the seasonal 500 hPa height as a function of the number of forecasts m. They have been obtained by the following procedure:

(1) for each year (out of the 16) drawing one member out of the 10 for the verification, the other 9 being averaged to produce the forecast;

(2) selecting at random m years for the forecasts and another m years for the verification (double scrambling);

(3) computing the C-score of these m forecast-verification pairs;

(4) repeating 500 times steps 1-2-3 (to obtain a sufficient stability of the results);

(5) ranking the 500 C-scores to extract the 25th and the 475th (5% and 95% quantiles of the distribution).

The relative symmetry of the boundaries of the confidence intervals is another indication of the absence of systematic bias. One can see that with $m=2$ (comparing the differences between two SST forcings as in Brankovic and Palmer, 1996) the C-score may be as high as 0.6 without any skill from the model.

Table 2 shows the C-scores for the 3 periods and the 3 fields. The scores are significantly positive: the 9 values are greater than the corresponding 95th percentiles (calculated as above by scrambling). The behavior of the midlatitude height is similar to that of European temperature, the

Table 1. *Empirical percentiles of the C-score distribution as a function of the number of forecasts (even values only) for a prediction of seasonal mean 500 hPa height over the northern hemisphere*

Forecasts	5%	95%
2	−0.57	0.59
4	−0.38	0.37
6	−0.32	0.30
8	−0.26	0.24
10	−0.24	0.20
12	−0.21	0.20
14	−0.19	0.18
16	−0.16	0.17

Forecasts and verifications have been scrambled so that the scores do not correspond to any skill.

Table 2. *C-scores for the 3 periods of midlatitude 500 hPa height (Z500), European domain 850 hPa temperature (T850), and tropical precipitation (Prec) for the 16 winter perfect model experiment*

	Week	Month	Season
Z500	0.81	0.34	0.31
T850	0.76	0.31	0.23
Prec	0.70	0.66	0.79

former being slightly more predictable than the latter. The tropical rainfall is quite different: the skill is higher and does not decrease with the time lag (the seasonal forecast is better than the monthly one). This can be easily explained by the fact that the main source of predictability is the tropical SST variability. The tropical precipitation, at least for the monthly or seasonal range, exhibits a significant response to the SST anomaly patterns imposed by the boundary conditions. The high scores for the week (days 7–14) are due to the fact that the initial error is weak and the model has no systematic error. They probably do not reflect the behavior of an actual forecast.

4. The scores versus the ensemble size

In order to investigate to what extent a perfect model forecast can be improved by increasing the size of the forecast ensemble, we have calculated C-scores for which one integration is drawn among the 10 members to make the verification, and p ($p \leqslant 9$) are drawn to be the forecast ensemble. For each year, the 10 members have been scrambled, and then the first is taken as the verification, and the next p as the forecasts. One can thus obtain a large number of C-scores. 500 such scores have been calculated and aggregated according to the rule defined in Section 3. Fig. 1 (solid line) shows the mean C-score of the 500 hPa height as a function of p. The increase does not seem to saturate with $p \leqslant 9$. It would be interesting to estimate the asymptotic value, and a size which gives a value close to this asymptote. The statistical calculation for this purpose is developed in Appendix A. If σ^2_{intra} is the intra-ensemble variance and σ^2_{inter} the interannual variance, then the estimated C-score is:

$$\hat{C} = \frac{\sigma^2_{\text{inter}} - \sigma^2_{\text{intra}}}{\sqrt{\sigma^2_{\text{inter}}(\sigma^2_{\text{inter}} - (1-1/p)\sigma^2_{\text{intra}})}}. \tag{10}$$

Introducing the ratio $r = \sigma^2_{\text{intra}}/\sigma^2_{\text{inter}}$, the asymptotic value is $C_\infty = \sqrt{1-r}$. Setting $p_0 = r/(1-r)$, eq. (10) can be written as:

$$\hat{C} = C_\infty \sqrt{\frac{p}{p+p_0}}, \tag{11}$$

where p_0 can be interpreted as the size of ensemble for which the C-score is $1/\sqrt{2}$ its asymptote. The squared correlation being interpreted in linear regression as the fraction of explained variance,

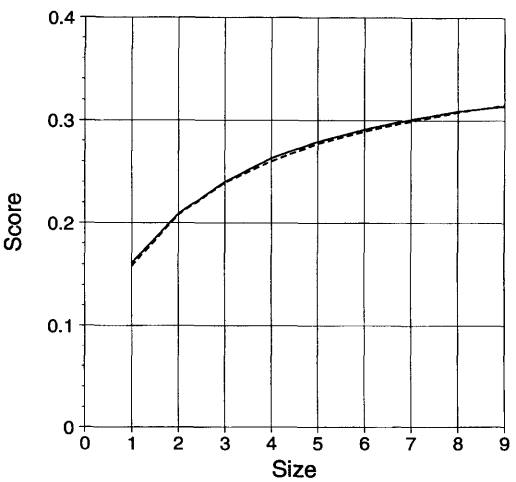

Fig. 1. Estimates of the C-scores of seasonal midlatitude 500 hPa height as a function of the ensemble size: random sub-sampling (solid line) and statistical estimation with the empirical momenta (dash line).

or as the maximum value for the SS (eq. (9)), taking $p=p_0$ corresponds to halving the asymptotic squared C-score.

Another interesting consequence of eq. (10) is that the asymptotic value C_∞ is the square root of the C-score for $p=1\hat{C}_1$: with a large number of twin integrations (one for the forecast, one for the verification), one can deduce easily the score improvement brought by a large ensemble. As expected, such an improvement is weak when the C-score is close to 0 or 1. The generalized relationship between $\hat{C}$ and $\hat{C}_1$ reads:

$$\hat{C}=\frac{\hat{C}_1}{\sqrt{1/p+(1-1/p)\hat{C}_1}}. \qquad (12)$$

A simple verification of the accuracy of the calculations of AppendixA is given by Fig. 1: one can see that the estimates with the Monte Carlo method are very close to those of the empirical momenta method.

Table 3 displays the values of C_∞ and p_0 for the 3 fields and the 3 periods. The behaviors of European temperature and midlatitude height are similar, except for seasonal forecasts. The European temperature fields require a larger ensemble to reach the asymptotic value. A value of 8 for p_0 does not mean that a good ensemble size should be 8, since the C-score is then 0.7 C_∞. The small value for p_0 in weekly forecasts is explained by the small initial error, and it is not guaranteed that a sample of 4 or 5 could be sufficient with actual forecasts. Moreover, only the impact on the score of the ensemble mean is studied here. A larger size could be necessary to estimate the ensemble spread. As far as the tropical precipitation is concerned, the scores remain high for the monthly and seasonal period (as shown in the last section), and the suitable ensemble size is less than 5. One can relate the condition on the ensemble size to a requirement for the ratio r between the spread and the interannual variability. If one wants that an ensemble size of 9 is greater than p_0, r must be less than 0.9. This is the case for our seasonal forecasts since r is respectively 0.83, 0.88, and 0.33 for Z500, T850 and precipitation. If one wants to reach a C-score of 0.9 times the asymptotic value, the size p must be approximately $5p_0$. In our case, the size is roughly 3 for precipitation, 20 for Z500, and 40 for T850.

Table 3. *Empirical parameters for the C-score variation with size (asymptotic value and characteristic size) for the 3 periods of midlatitude 500 hPa height (Z500), European domain 850 hPa temperature (T850), and tropical precipitation (Prec)*

	Week C_∞/p_0	Month C_∞/p_0	Season C_∞/p_0
Z500	0.84/0.5	0.42/5	0.40/5
T850	0.79/0.5	0.40/5	0.33/8
Prec	0.73/1	0.70/1	0.81/0.5

The use of the perfect model approach also enables study of the geographical distribution of the scores. Figs. 2–4 show the asymptotic value of the local C-scores (i.e., $C_\infty(x)$) for the 500 hPa height, 850 hPa temperature, and precipitation fields. The local C-score is nothing but the temporal correlation (calculated for the 16 years) of the seasonal forecast and verification series. All the fields have their maximum in the tropics. For Z500, this maximum is longitude independent. For precipitation, and to a lesser extent for temperature, this maximum is located over the ocean. The precipitation, and to a lesser extent the temperature and height, have a secondary maximum at high latitudes. This predictability could be produced by the interannual variability of sea-ice extents which may cause large surface temperature anomalies to occur. In the northern midlatitudes correlations below 0.2 are observed over Europe, California and North-eastern United States. These areas therefore show little sensitivity to the SST variations of the 1979–1994 period with our model. Further studies with other models, other SST anomalies, and actual forecast conditions will confirm or refute that these area correspond to minima of predictability.

The year-to-year distribution of the C-score can also be investigated. One must keep in mind that the label of the years (e.g., 1983) refers to the SST forcing only, since we are dealing with virtual years with a perfect model. Since we restrict here to seasonal averages (January-February-March), this label will correspond to the year of the January month, not to the year of the starting situation. It can be shown that eq. (12) is still valid for one forecast (the calculations of Section 7 concern the average for m forecasts). Figs. 5–7 display the C-scores of the 3 fields for each winter, and an increasing ensemble size (1, 3, 9 and ∞). As far as the midlatitude geopotential is concerned, 1989 is the best year with an

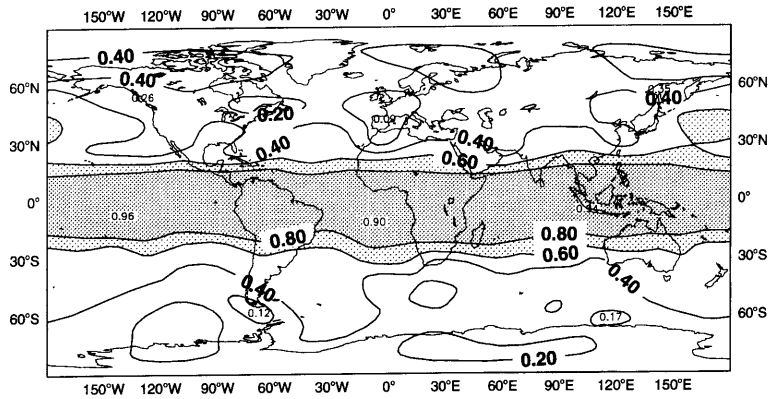

Fig. 2. 500 hPa height *C*-score map (i.e., time correlation) estimated from the perfect model for an infinite ensemble size; contour interval 0.20, shading above 0.60.

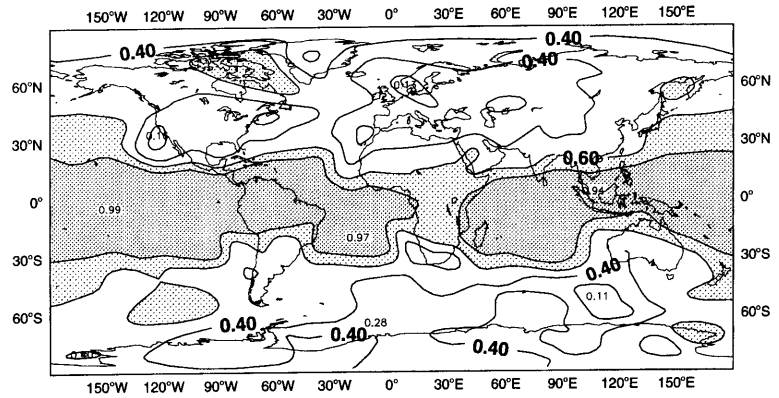

Fig. 3. As Fig. 2 for 850 hPa temperature.

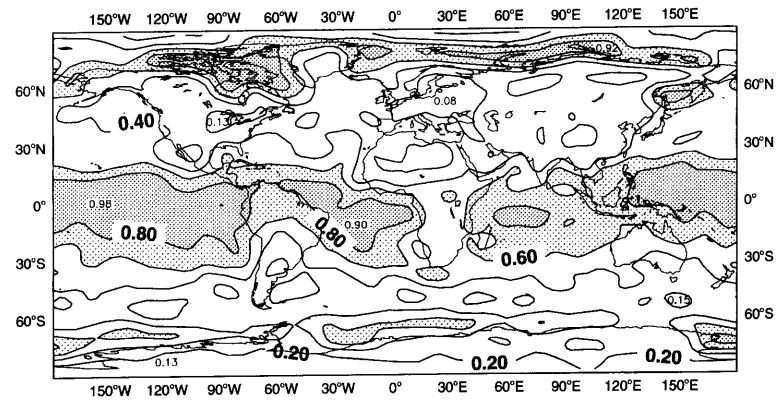

Fig. 4. As Fig. 2 for precipitation.

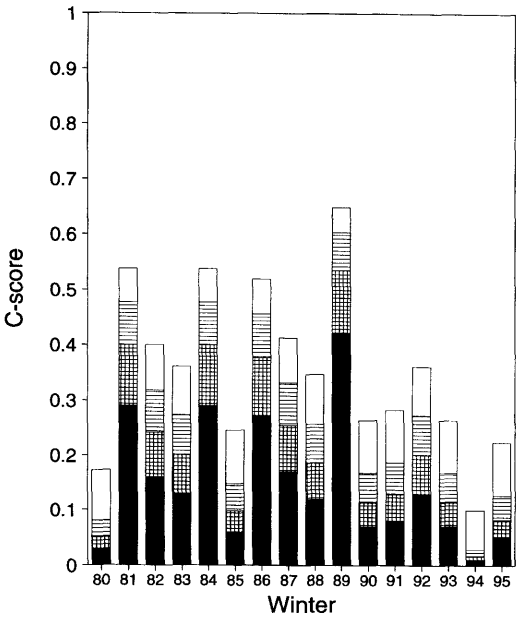

Fig. 5. Estimates of the *C*-scores of seasonal midlatitude 500 hPa height as a function of the ensemble size for each winter (JFM): 1-member ensemble (black filling), 3-member ensemble (vertical hatching), 9-member ensemble (horizontal hatching), and asymptote (no filling).

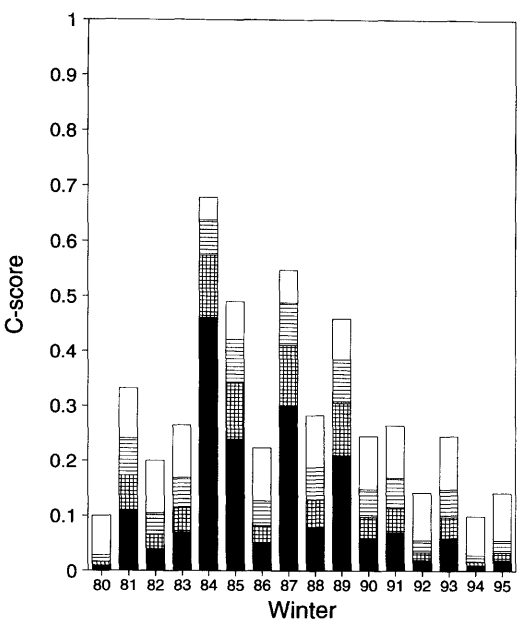

Fig. 6. As Fig. 5 for 850 hPa temperature over Europe.

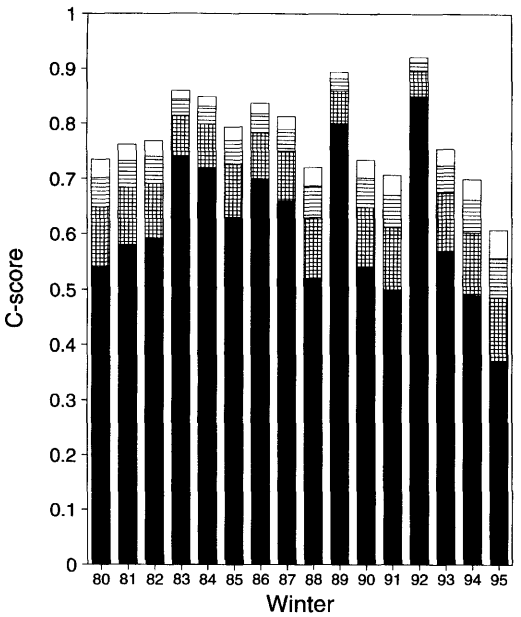

Fig. 7. As Fig. 5 for tropical precipitation.

asymptotic score greater than 0.60. This winter corresponds to the end of a strong negative ENSO event. The positive ENSO winters (1983 and 1987) are in the average: the impact of the SST anomalies is not dominant in the model midlatitude circulation. For European temperature, 1987 is the second best winter, after 1984. For tropical precipitation, 1983 is the 3rd best out of the 16 winters. Assuming that the interannual variability of our estimates is significant, this shows that the impact of SST anomalies on the model response is not strongly correlated with their amplitude. Fig. 7 illustrates that in the tropics a 1-member forecast is not widely improved by ensemble averaging. The JFM 1983 500 hPa height anomalies (not shown) are similar to the observed ones only in the Pacific (increase along the equator and decrease at midlatitudes). The classical Pacific North America (PNA) pattern is simulated with the right sign, but with an amplitude lower than 10 m against more than 50 m in the observations. In 1989 (Fig. 8), the agreement between the simulation (a) and the observation (b) is better than in 1983. Both figures show an amplification of the zonal circulation in the Atlantic with a deepening of the Icelandic low and a blocking over the Pacific.

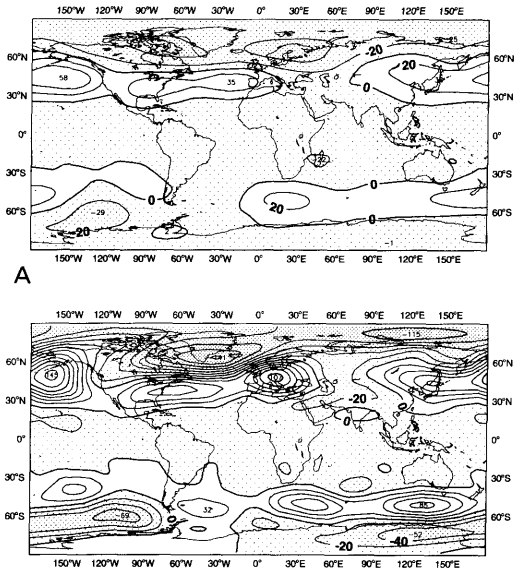

Fig. 8. 500 hPa height anomaly for JFM 1989 in the perfect model experiment (a) and in the observations (b): contour interval 20 m, shading over the negative anomalies.

The pattern in the southern hemisphere is also well captured. This (best) case shows that:

(i) the pure response to SST anomalies may agree with observed simulated anomalies;

(ii) this occurs when the intra-ensemble agreement is high

The lower amplitude of the predicted pattern is a combination of ensemble averaging and underestimated interannual variability by the model.

5. Conclusion and perspectives

The predictability of various fields, areas, and time ranges has been studied using the "perfect model" approach. The main source of predictability in this experiment comes from the SST variability of the 1979–1994 period. The mean scores of winter forecasts vary from large values for the tropical precipitation to weak values in the midlatitudes, in particular over Europe. Empirical estimates of the scores with the size of the forecast ensemble have shown that the size necessary to approach the asymptotic value are much larger for temperature over Europe than for tropical precipitation, in the day 30–120 range.

These results are not surprising, since most of the interannual SST variability is located in the tropics. The most important variability is in the Pacific ocean (ENSO). ENSO's impact on remote regions like Europe exists, but explains a weak part of the interannual variability (Fraedrich, 1994).

The results of our study give a more quantitative evaluation to well known features of the long-range predictability. They depend however on the perfect model approach and on the model itself. The results from the IAPO experiment will give more reliable estimates, since the statistical method can be generalized to an imperfect model. The impact of ensemble size on monthly mean imperfect model scores has been studied in Déqué (1991), and it was shown that a size of 5 members was sufficient for Northern Hemisphere 850 hPa temperature. The equivalent of C_∞ in that study was 0.27 (here it is 0.40 from Table 3).

The scores are not very encouraging for the midlatitude seasonal forecasts: correlation of 0.40 with a perfect model and an infinite ensemble. It is costly to have a large ensemble, and even more difficult to have a quasi-perfect model. However, this "perfect model" is perfectible. Indeed, the *C*-score is a decreasing function of the ratio r of the intra-ensemble variance (the model mean spread) over the interannual variance. The ability of the model to reproduce the intra-ensemble variance and its year-to-year variability is unknown, but the underestimation of the interannual variance by the model is well identified: using the first 12 yeas of the IAPO experiment which are available, and the first 12 years of PAPO, the interannual standard deviation of midlatitude seasonal 500 hPa height is 43 m (observed) against 31 m (simulated). In the perfect model experiment, the amplitude of the anomalies in the slowly evolving initial conditions (e.g., deep soil moisture or QBO) are thus underestimated. For example, the model stimulates a realistic QBO, except that its amplitude is 5 times smaller than observed (Cariolle et al., 1993). Thus the predictability potential of the initial conditions, which could be as important as the SST in the midlatitudes, is not fully exploited by the perfect model.

In order to evaluate the "imperfection" of the perfect model approach, the seasonal midlatitude 500 hPa height scores have been compared for the 12 winters (JFM) 1980–1991 of PAPO and IAPO experiments. The perfect model yields a *C*-score of 0.36 (slightly greater than the estimate

with the 16 winters) and the IAPO experiment yields a value of 0.32. A "pseudo-perfect model" can be obtained by taking one of the forecasts of IAPO as the observation and the other 8 as the forecast ensemble. This approach better takes into account the information brought by the variability of the initial conditions. As expected, the score is larger (0.46). These results indicate that the perfect model approach does not strongly overestimate the actual skill, at least for midlatitude Z500. Indeed, the preliminary scores for T850 over Europe are 0.11 (IAPO) against 0.31 (PAPO), and the scores for tropical precipitation are 0.30 (IAPO) against 0.78 (PAPO).

The year-to-year variability of the scores is greater in IAPO than in PAPO. As far as midlatitude Z500 is concerned the extremes are -0.31 in JFM 1991 and 0.74 in 1989 (IAPO) and 0.05 in 1980 and 0.60 in 1989 (PAPO, see Fig. 5). It is interesting to notice that 1989 is in both cases the best year. The strongest SST forcing occurred in 1983. This is the second best year of IAPO, but a normal year for PAPO. As far as tropical precipitation is concerned, the scores range between 0.0 in 1981 and 0.53 in 1983 for IAPO, and between 0.70 in 1981 and 0.88 in 1989 for UAPO (1983 is the second best year with 0.86). As far as European temperature is concerned, the amplitude of the score variations is similar in IAPO and PAPO, without consistency between the two series. This enhanced variability in IAPO scores reveals another limitation of the perfect model approach.

A simple overestimate of the Z500 score can be obtained by assuming that improving the model would leave the intra-ensemble variance unchanged, and would double the interannual variance in order to reach the observed value. Then, the calculations of Section 4 indicate that the asymptotic C-score of 0.40 would become 0.76. If such a score is reached in the future, then the numerical seasonal forecast in the midlatitude will be operationally of great use. With smaller scores, the main fraction of the signal will remain unpredictable, but the small predictable part may bring some benefit against a random strategy when the results are averaged over a long period.

6. Acknowledgements

The author is indebted to his colleagues from Météo-France and ECMWF who developed the numerical forecast model, and in particular to J. F. Geleyn. The ECMWF reanalyses were used in this study to provide boundary conditions to the simulations. Thanks are also due to P. Gleckler for his help in reading the manuscript, and to J. Ph. Piedelievre for running the IAPO experiment. This work was partly supported by the Commission of European Union (contracts AVI-CT93-0010 and ENV4-CT95-0109), by PNEDC (French National Program for the Study of Climate Dynamics), and by the Regional Council of Midi-Pyrénées.

7. Appendix

Statistical estimation of the C-score

In this appendix, we want to estimate the C-score of a perfect model, using only the statistical properties of the distribution of a field X at a given location. The following statistical problem is considered: for each year i ($i=1, \ldots, m$ with here $m=16$), $p+1$ values ($X_{i0}, X_{i1}, \ldots, X_{ip}$) are drawn independently with the same probabilistic law which depends on i. This law is unknown, but may be estimated with a sample of n independent equiprobable values ($F_{i1}, F_{i2}, \ldots, F_{in}$). Here, $n=10$, because the reference and the 9 forecasts have the same probability distribution. The C-score is calculated from 3 empirical covariances as:

$$C = \frac{C_{VF}}{\sqrt{C_{VV} C_{FF}}}, \tag{13}$$

where F is the ensemble mean forecast and V is the verification. We do not take directly C as an estimator of the C-score, but estimate separately the 3 covariances, and set:

$$\hat{C} = \frac{\hat{C}_{VF}}{\sqrt{\hat{C}_{VV} \hat{C}_{FF}}}. \tag{14}$$

Let us first estimate for a given year i the 3 covariances. The forecast is the average of the random variables ($X_{i1}, \ldots, X_{ip}$) and the verification is X_{i0}. The 3 empirical momenta are:

$$C_{FV}(i) = \left(\frac{1}{p} \sum_{j=1}^{p} X_{ij} - \frac{1}{mp} \sum_{j=1}^{p} \sum_{k=1}^{m} X_{kj} \right) \times \left(X_{i0} - \frac{1}{m} \sum_{k=1}^{m} X_{k0} \right), \tag{15}$$

$$C_{FF}(i)=\left(\frac{1}{p}\sum_{j=1}^{p}X_{ij}-\frac{1}{mp}\sum_{j=1}^{p}\sum_{k=1}^{m}X_{kj}\right)^2, \qquad (16)$$

$$C_{VV}(i)=\left(X_{i0}-\frac{1}{m}\sum_{k=1}^{m}X_{k0}\right)^2 \qquad (17)$$

Now we have to calculate the statistical expectation of the above 3 quantities, because X_{ij} is a random variable. Since we have assumed that a member of an ensemble was independent of the other members of the ensemble, and a fortiori of the members of other ensembles, we can write $E(X_{kj}X_{k'j'})=E(X_{kj})E(X_{k'j'})$ when $j\neq j'$ or when $k\neq k'$. The fact that the different members of an ensemble are independent does not contradict the fact that they have the same statistical properties, in particular the same mean.

If we expand the polynomials in the right hand members of the above 3 equations, and take the expectation of both members, using the linearity of $E(\ )$, and the fact that the momenta of X_{kj} do not depend on j, and thus are the same as those of X_{k0}, we obtain:

$$E(C_{FV}(i))=E(X_{i0})^2-\frac{2}{m}\sum_{k=1}^{m}E(X_{k0})E(X_{i0})+\frac{1}{m^2}\sum_{k=1}^{m}\sum_{k=1}^{m}E(X_{k0})E(X_{k'0}), \qquad (18)$$

$$E(C_{FF}(i))=\frac{1}{p}E(X_{i0}^2)+\frac{p-1}{p}E(X_{i0})^2-\frac{2}{mp}E(X_{i0}^2)-\frac{2}{mp}\sum_{k\neq i}E(X_{k0})E(X_{i0})-\frac{2(p-1)}{mp}\sum_{k=1}^{m}E(X_{k0})E(X_{i0})+\frac{1}{m^2p}\sum_{k=1}^{m}E(X_{k0}^2)+\frac{1}{m^2p}\sum_{k=1}^{m}\sum_{k'\neq k}E(X_{k0})E(X_{k'0})+\frac{p-1}{m^2p}\sum_{k=1}^{m}\sum_{k'=1}^{m}E(X_{k0})E(X_{k'0}). \qquad (19)$$

$$E(C_{VV}(i))=E(X_{i0}^2)-\frac{2}{m}E(X_{i0}^2)-\frac{2}{m}\sum_{k\neq i}E(X_{k0})E(X_{i0})+\frac{1}{m^2}\sum_{k=1}^{m}E(X_{k0}^2)+\frac{1}{m^2}\sum_{k=1}^{m}\sum_{k'=\neq k}E(X_{k0})E(X_{k'0}). \qquad (20)$$

If we sum the above 3 equations for $i=1$ to m (i is the index of the year) and divide by m, we get the expectation of the time averages. Omitting the subscript 0, some simplifications lead to:

$$E(C_{FV})=\frac{m-1}{m^2}\sum_{i}(E(X_i))^2-\frac{1}{m^2}\sum_{i\neq k}E(X_i)E(X_k), \qquad (21)$$

$$E(C_{FF})=\frac{m-1}{m^2}\sum_{i}\left(\frac{p-1}{p}E(X_i)^2+\frac{1}{p}E(X_i^2)\right)-\frac{1}{m^2}\sum_{i\neq k}E(X_i)E(X_k), \qquad (22)$$

$$E(C_{VV})=\frac{m-1}{m^2}\sum_{i}E(X_i^2)-\frac{1}{m^2}\sum_{i\neq k}E(X_i)E(X_k). \qquad (23)$$

The expectations involved in the above equations have to be estimated. For that, we use our n-samples (F_{ij}). Setting the overbar for the ensemble average (i.e., the average with respect to j), we estimated $E(X_i^2)$ by $\overline{F_i^2}$ and $E(X_i)E(X_k)$ by $\overline{F_iF_k}$ when $k\neq i$. One must take care that $\overline{F_i^2}$ is a biased estimate of $E(X_i)^2$; a simple unbiased estimate is $(n\overline{F_i^2}-\overline{F_i^2})/(n-1)$.

Once the 3 covariances are estimated as a function of p, they yield, through eq. (14), an estimate of the C-score. If we express the C-score as a function of the momenta $E(X_i)$ and $E(X_i^2)$, the formula is long and difficult to interpret. Let

$$\sigma^2_{\text{intra}}=\frac{1}{m}\sum_{i}(E(X_i^2)-E(X_i)^2), \qquad (25)$$

$$\sigma^2_{\text{inter}}=\frac{m}{m-1}\left\{\frac{1}{m}\sum_{i}X_i^2-\left(\frac{1}{m}\sum_{i}X_i\right)^2\right\}, \qquad (26)$$

be the intra-ensemble and the interannual variances of the meteorological field X. Then, if we expand σ^2_{inter} and eliminate $E(X_i)$ and $E(X_i^2)$ in the equations, the 3 covariances may be rewritten as:

$$E(C_{FV}) = \frac{m-1}{m}(\sigma^2_{\text{inter}} - \sigma^2_{\text{intra}}), \tag{27}$$

$$E(C_{FF}) = \frac{m-1}{m}(\sigma^2_{\text{inter}} - \frac{p-1}{p}\sigma^2_{\text{intra}}), \tag{28}$$

$$E(C_{VV}) = \frac{m-1}{m}\sigma^2_{\text{inter}}. \tag{29}$$

and the estimated C-score may be written as:

$$\hat{C} = \frac{\sigma^2_{\text{inter}} - \sigma^2_{\text{intra}}}{\sqrt{\sigma^2_{\text{inter}}(\sigma^2_{\text{inter}} - (1-1/p)\sigma^2_{\text{intra}})}}. \tag{30}$$

This equation is valid for individual grid points, and also for spatial averages, provided that both variances are averaged over the domain before entering into the formula.

REFERENCES

Barker, T. W. 1991. The relationship between spread and forecast error in extended-range forecasts. *J. Climate* **4**, 733–742.

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

Brankovic, C. and Palmer, T. N. 1996. Atmospheric seasonal predictability and estimates of ensemble size. *Mon. Wea. Rev.*, in press.

Brankovic, C., Palmer, T. N., Molenti, 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.

Cariolle, D. and Déqué, M. 1986. Southern hemisphere medium-scale waves and total ozone disturbances in a spectral general circulation model. *J. Geophys. Res.* **91**, 10825–10846.

Cariolle, D., Amodei, M., Déqué, M., Mahfouf, J F., Simon P. and Teyssèdre, H. 1993. A quasi-biennial oscillation signal in general circulation model simulations. *Science* **261**, 1313–1316.

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.

Déqué, M., Royer J F., and Stroe, R., 1994a. Formulation of gaussian probabilty forecast based on model extended-range integrations. *Tellus* **46A**, 52–65.

Déqué, M., Dreveton, C., Braun A. and Cariolle, D. 1994b. The ARPEGE/IFS atmosphere model: a contribution to the French community climate modelling. *Climate Dyn.* **10**, 249–266.

Fraedrich, K. 1994. An ENSO impact on Europe? A review. *Tellus* **46A**, 541–552.

Geleyn, J. F. and Hollingsworth, A. 1979. An economical analytic method for the computation of the interaction between scattering and line absorption of radiation. *Beit. Phys. Atmos.* **52**, 1–16.

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

Houtekamer, M. and Derome, J. 1994. Prediction experiment with two-member ensembles. *Mon. Wea. Rev.* **122**, 2179–2191.

Louis, J. F. 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary Layer Meteorology* **17**, 187–202.

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

Noihlan, J. and Planton, S. 1989. A simple parameterization of land surface processes for meteorological models. *Mon. Wea. Rev.* **117**, 536–549.

Palmer, T. N. and Anderson, D. L. T. 1994. The prospect for seasonal forecasting — a review paper. *Quart. J. Roy Meteor. Soc.* **120**, 755–793.

Royer, J. F. 1993. Review of recent advances in dynamical extended-range forecasting for the extratropics. In: *Proceedings of the NATO Workshop on Prediction of interannual climate variations.* Nato ASI Series Publication. Global environmental change, vol. 16, pp. 49–69, Springer-Verlag.

Savijärvi, H. 1995. Error growth in a large numerical forecast system. *Mon. Wea. Rev.* **123**, 212–221.

Stern W. and Miyakoda, K. 1995. Feasibility of seasonal forecasts inferred from multiple GCM simulations. *J. Climate* **8**, 1071–1085.

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.